United Nations
Interregional Crime and Justice
Research Institute

THINKING IN ALTERNATIVES

A Community Note for Southeast Asia: Strengthening Responses to Terrorist and Violent Extremist Use of the Internet

UNICRI PUBLICATIONS

THINKING IN ALTERNATIVES

A Community Note for Southeast Asia: Strengthening Responses to Terrorist and Violent Extremist Use of the Internet

United Nations
Interregional Crime and Justice
Research Institute

DISCLAIMER

ACKNOWLEDGEMENTS

This report by the United Nations Interregional Crime and Justice Research Institute (UNICRI) was written by Arthur Bradley, UNICRI Consultant, under the overall guidance and editing of Ottavia Galuzzi (UNICRI) and Odhran McCarthy (UNICRI), with editing by Christopher Plagnol (UNV) and design by Marianna Fassio (UNICRI).

UNICRI expresses gratitude to the Ministry of Home Affairs of Singapore for the fruitful collaboration in hosting the regional workshop titled *Thinking in Alternatives: Exploring Open-Source Intelligence (OSINT) and Community-Driven Approaches to Content Moderation for Terrorist and Violent Extremist Content Online*, held on 2-3 December 2025 in Singapore. In addition, UNICRI expresses gratitude to all participants in the workshop for their active participation, expertise, and insights. Particular thanks and appreciation are extended to colleagues at the United Nations Office of Counter-Terrorism (UNOCT) for their support throughout the workshop and for providing feedback on this report.

The use of the Internet and online ecosystems by terrorist and violent extremist actors has become increasingly complex, intersecting with a broad range of harmful narratives, violent content and online subcultures. These risks are compounded by limited content moderation capacities; a lack of awareness among users and some technology companies of what constitutes harmful and unlawful content; linguistic and cultural gaps; and weaker partnerships between governments, civil society, and information and communication technology companies. The situation has been exacerbated in recent years by the global rollback of content moderation policies, as well as capacity and enforcement issues faced by some major social media platforms. Collectively, these developments have reduced safeguards against harmful content in general and created a more permissive environment for terrorist and violent extremist narratives.

In this complex and evolving digital landscape, responses to terrorism and violent extremism online must adapt and evolve accordingly. Existing approaches and solutions often prove to be limited in their effectiveness and adaptability, despite their strengths. Alternative ways of thinking and a more holistic approach are needed to encourage cross-learning among diverse sectors of information and communication technology companies, law enforcement and civil society. Alternative approaches include leveraging open-source intelligence (OSINT), sharing information on similar behavioural signals relevant to terrorism, violent extremism, cybercrime and other online harms, and collaborating meaningfully at the national level and within Member States. Countering terrorism and violent extremism online also requires effective content moderation rooted in a deep understanding of linguistic and cultural norms, while upholding human rights. To achieve this, diverse actors should be empowered – primarily civil society organizations and technology hubs – to co-create community-driven solutions that improve content moderation at the local, national and regional levels. To this end, this report explores alternative approaches to detecting, monitoring and responding to terrorist and violent extremist use of the Internet and online ecosystems, while encouraging readers to think beyond what is already known and consider if and how solutions can be adapted to specific contexts.

Although online threats, methods and approaches are global in nature, this report focuses on Southeast Asia to examine the threats in that region and to explore how approaches and solutions can be adapted to specific contexts. This regional focus draws on the findings of the regional workshop *Thinking in Alternatives: Exploring Open-Source Intelligence (OSINT) and Community-Driven Approaches to Content Moderation for Terrorist and Violent Extremist Content Online*, convened by UNICRI and the Ministry of Home Affairs of Singapore on 2-3 December 2025 in Singapore. This report complements the findings of the regional workshop through consultation with international and national experts and desk-based research.

The report provides an overview of the current terrorist and violent extremist threat landscape in Southeast Asia and examines how terrorist and violent extremist actors exploit digital technologies in the region to advance their objectives. Building on this analysis, it explores complementary and alternative approaches to strengthen responses to terrorist and violent extremist content online. On detection and monitoring, it examines the use of OSINT methods to identify and track these threats, including limitations, challenges, and areas for improvement. On the response side, it explores content moderation, approaches, with particular attention to community-based mechanisms such as user reporting, community notes, middleware, open-source tools, and moderation on the so-called fediverse – a network of independent social media servers that allow users to interact on different platforms without being controlled by a single company.

This work builds on UNICRI's *Cybercrime and Online Harms Workstream*, which seeks to bridge the gap between complex, technical threats and practical, human-centred approaches, ensuring that cyberspace remains safe and secure and not a vector for exploitation by malicious actors. It also builds on longstanding cooperation between UNICRI and the United Nations Office of Counter-Terrorism (UNOCT) in the domains of cybersecurity and new technologies, in particular in addressing terrorist and violent extremist exploitation of digital technologies.[1]

While this report does not cover the relevant distinction between terrorism and violent extremism, particularly their manifestations online, it does recognize the importance of such distinction and how it should be reflected in the approaches explored in this report. This distinction is particularly important in the context of content moderation policies, as frameworks developed to address terrorist content cannot be indiscriminately applied to violent extremism. In practice, regulations targeting violent extremism have at times been misused to restrict legitimate dissent, resulting in adverse impacts on freedom of expression and human rights more generally.[2]

Current Threat Landscape in Southeast Asia

Southeast Asia faces significant threats to its security from terrorism and violent extremism – exacerbated by the growing, steady and multi-faceted online presence across its populations. In fact, use of digital technologies in Southeast Asia is often higher compared to other regions, with more than 60% of its population being active social media users.[3] Furthermore, those social media users use an average of at least seven platforms, exceeding the global average. In this context, the Philippines, Thailand and Indonesia are in the top five countries globally for the consumption of video content online.[4] In addition, online gaming and esports markets continue to grow: more than 80% of the region's urban population were players of online games in 2024, and the industry's 2025 revenue is forecasted at USD$6.39 billion.[5] Digital technology usage statistics in Southeast Asia indicate a broad potential recruitment pool for terrorist and violent extremist actors.

The affordances of digital technologies significantly shape the terrorism threat landscape in Southeast Asia. Terrorist and extremist networks maintain a persistent presence on an increasingly broad range of platforms and services, including social media, messaging apps, online gaming and game-adjacent platforms, file-sharing services, video-sharing platforms, and static websites. These actors use digital technologies widely for propaganda dissemination, recruitment, financing and internal communications. At the same time, algorithmic recommendation systems increase the risk that users encounter such materials unintentionally. Consequently, user risk is exacerbated by several factors, including the rise in local online media outlets, the spread of grey-area content in mainstream social media platforms, and the translation of harmful content into local languages.

1 UNICRI, UNCCT, Beneath the Surface: Terrorist and Violent Extremist Use of the Dark Web and Cybercrime-as-a-Service, June 2024; UNICRI, VOX-Pol, Right- and Left-Wing Violent Extremist Abuse of Digital Technologies in South America, Africa and Asia, April 2025; UNICRI, UNCCT, New Quest Unlocked: The Intersection of Gaming and Violent Extremism, December 2025.

2 United Nations, General Assembly, Human Rights Council, Impact of Measures to Address Terrorism and Violent Extremism on Civic Space and the Rights of Civil Society Actors and Human Rights Defenders, A/HRC/40/52, March 2019.

3 Sue Howe, Social Media Trends and Habits in Southeast Asia, Meltwater, January 2025.

4 Adam Shaw, Digital 2025 April Statshot by We Are Social: Southeast Asians Among the World's Highest Short-form Video Consumers, Campaign Brief Asia, April 2025.

5 The Gaming Market in Southeast Asia, Allcorrect Games, February 2025.

The intent of global terrorist groups – such as the Islamic State in Iraq and the Levant (ISIL/Da'esh) and Al-Qaida – to conduct or inspire attacks in the region is persistent, as noted in the most recent annual threat assessment published by Singapore's Ministry of Home Affairs.[6] The broader related violent Islamist ideology continues to drive radicalization. While organized groups continue to operate in the region, such as in the southern Philippines, the threat posed by groups affiliated with terrorist organizations (such as ISIL and Al-Qaida) increasingly emanate from lone actors and small cells, making identification, monitoring and intervention against offenders more challenging. In these cases, not only is it more difficult to identify linked behaviour and user groups, but detection requires greater use of vectors, such as external URLs and linked usage of content on other online platforms.

In Southeast Asia, violent extremism associated with right-wing ideologies is gaining traction, along with ethnonationalist movements and references to Austronesian supremacy. Singapore's threat assessment identifies far-right extremism as an emerging threat in which domestic adherents have expressed a strong antipathy towards Malay and Muslim communities in Singapore. Other forms of so-called right-wing extremism are present across other national contexts in Southeast Asia.[7]

According to Singapore's Ministry of Home Affairs, Singapore identified and handled four youths subscribing to violent extremist ideologies between 2020 and June 2025. Digital networks of individuals such as these represent diversity in ideologies – reflecting the diversity of populations across the region – and can include elements of so-called East-Asian supremacy, Muslim nationalism, anti-Rohingya or anti-Muslim prejudice, Buddhist ultranationalism, antisemitism, and support for authoritarianism.[8] A so-called glocalization effect (a portmanteau of globalization and localization) is apparent in such cases, that is, when local and global violent extremist grievances and trends are merged. These developments reflect broader manifestations of terrorism and violent extremism motivated by xenophobia, racism and other forms of intolerance.

Users in Southeast Asia are highly exposed to a variety of emerging violent and coercive online networks. During the 2025 regional workshop in Singapore, several experts referred to growing threats they described as 'nihilistic violent extremism', as well as to online communities such as the True Crime Community,[9] and highlighted how these intersect with elements from violent extremist ideologies and the collective adulation of mass killers.[10] In Indonesia, the National Counter-Terrorism Agency (BNPT) affirmed that the perpetrator of the bombing at SMAN 72 high school in Jakarta in November 2025 was inspired in part by right-wing ideologies, including narratives espoused by the perpetrator of the attack on two mosques in Christchurch, New Zealand, in March 2019,[11] and had been actively

6 Singapore's Internal Security Department, Singapore Terrorism Threat Assessment Report 2025, July 2025.

7 It should be noted that there is no universally agreed definition of "far-right violent extremism". The term is used here in a descriptive sense, reflecting terminology commonly used by practitioners in the region and during the workshop. Within United Nations frameworks, acts motivated by ideologies commonly associated with such terminology are generally considered under the broader category of terrorism committed on the basis of xenophobia, racism and other forms of intolerance, or in the name of religion or belief, as reflected in the eighth review of the Global Counter-Terrorism Strategy (A/RES/77/298).

8 UNICRI, VOX-Pol, Right- and Left-Wing Violent Extremist Abuse of Digital Technologies in South America, Africa and Asia, April 2025.

9 Institute for Strategic Dialogue, "Memetic violence: How the True Crime Community generates its own killers", October 2025.

10 It should be noted that there is no universally agreed definition of the term "nihilistic violent extremism." The term is increasingly used in practitioner and policy circles to describe emerging forms of extreme violence associated with loosely structured online communities that lack coherent ideological foundations but encourage or glorify violence. However, the term does not fully capture the diversity of actors, motivations and harms associated with these phenomena, and there is no clear consensus on whether these forms of violence should be considered a form of violent extremism. In some contexts, including in parts of Europe, similar dynamics are instead described as forms of "nihilistic extremist violence." The use of the term in this document reflects the terminology adopted by participants during the workshop and should not be understood as implying a universally accepted definition or classification.

11 Ananda Teresia, "Suspect in Indonesia Mosque Bombing Was Inspired by Past Mass Killings, Police Say", Reuters, November 2025.

involved with the True Crime Community online.[12] Rather than applying the label 'nihilistic violent extremism', official sources[13] and experts described this type of attack – including the Siam Paragon mall shooter[14] – as memetic violence. The term reflects the fact that perpetrators demonstrate primarily performative emulation with aesthetic appropriation and minimal ideological coherence and commitment.[15] In reference to the SMAN 72 high school bombing, the National Counter-Terrorism Agency (BNPT) underscored this memetic element, noting that users (predominantly youth within these communities) tend to imitate previous acts of violence to gain notoriety.

Youth radicalization and minors' involvement in violent extremism are on the rise globally, and Southeast Asia is no exception. The annual threat assessment of Singapore's Ministry of Home Affairs notes the rise of far-right ideologies among youth, fuelled by the accessibility of extremist and terrorist content online, including on mainstream platforms.[16] In 2024, Indonesia's National Counter-Terrorism Agency (BNPT) recorded more than 180,000 pieces of extremist content online, the majority appearing on Instagram, followed by Facebook and TikTok.[17] Reported cases of children[18] across the region being exposed to violent extremist ideologies and content through coercive online networks are on the rise, emanating from the True Crime Community.

12 Indonesia National Police, "SMA 72 Bomber Linked to Online "True Crime Community: BNPT Chief", November 2025.

13 Wildan Noviansah, "Densus 88 Ungkap Fenomena 'Memetic Violence' di Ledakan SMAN 72 Jakarta," Detik News, November 2025.

14 Kocha Olarn, Helen Regan, "Teen Suspect in Fatal Thai Shopping Mall Shooting Charged with Murder, Police Say", CNN World, 4 October 2023.

15 Munira Mustaffa, "Cosplaying Columbine: How Memetic Violence Transformed Southeast Asia's Extremist Threat", GNET, December 2025.

16 Munira Mustaffa, "Radical Right Activities in Nusantara's Digital Landscape: A Snapshot", GNET, April 2022.

17 Muhammad Makmun Rasyid, "Southeast Asia Facing Hidden Extremist Threat", Bangkok Post, July 2025.

18 "Densus 88 Tangani 68 Anak Terpapar Ideologi Neo-Nazi di 2025, Sasarannya Sekolah", Kumparan News, December 2025.

Challenges in Monitoring and Moderating Terrorist Content in Southeast Asia

Law enforcement and other public-sector agencies face significant challenges in monitoring and responding to terrorist and violent exploitation of digital technologies in Southeast Asia. The scale of the threat is not insignificant. In 2023, Tech Against Terrorism found terrorist content on 242 platforms, the majority produced by global groups associated with ISIL/Da'esh and Al-Qaida.[19] The sheer range of platforms and services presents a serious challenge for authorities to identify where and when exploitation is taking place, and then to prioritize resources in response.

Investigators require specialist knowledge to safely and effectively navigate these expansive digital ecosystems, to identify relevant information, and to capture necessary evidence. This includes an understanding of a range of threat actors and technologies: the decentralized web, messaging apps with varying degrees of encryption, static websites, file-sharing platforms, video-sharing platforms, archiving services, and link shorteners. Given the evolving nature of these technologies, expertise must be continually maintained by regular primary source monitoring or training.[20]

Gaps continue to exist in cross-sector and cross-jurisdictional cooperation in the online counter-terrorism space in Southeast Asia. In seeking to improve safety online, governments and law enforcement agencies must exert influence on privately-owned companies that may not have a physical presence in their jurisdiction and are likely operated from abroad.[21] Finally, misalignments in definitions of contentious concepts like terrorism, violent extremism and hate speech compound these enforcement issues.[22]

Identifying and removing terrorist and violent extremist content originating in Southeast Asia presents difficulties even for the largest technology companies.[23] The region is among the most linguistically diverse in the world, challenging companies to correctly interpret cultural nuance and slang in online posts from specific local areas, particularly if they lack a sufficient number of employees fluent in relevant languages.[24] In fact, linguistic difficulties have likely increased due to widespread layoffs across the sector in recent years, including the dismissal of approximately 500 content moderators at TikTok in Malaysia in October 2024.[25]

19 Maura Conway, Ashley A. Mattheis, Sean McCafferty and Miraji H. Mohamed, Violent Extremism and Terrorism Online in 2023: The Year in Review, Policy Report, Publications Office of The European Union, 2023.

20 UNCCT, UNICRI, Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and Southeast Asia, 2021.

21 Beltsazar Krisetya, "From Principles to Protocols: Embedding Partnerships into Content Moderation Technologies against Mis/Disinformation", ISEAS, September 2025.

22 "Malaysia Warns TikTok, Meta Over Alleged Blocking of Pro-Palestinian Content", Reuters, October 2023.

23 Nuurrianti Jalli, "Holding Social Media Companies Accountable for Enabling Hate and Disinformation", Yusof Ishak Institute, July 2024.

24 "How Big Tech Platforms Are Neglecting Their Non-English Language Users", Global Witness, November 2023.

25 Rozanna Latiff, "ByteDance's TikTok Cuts Hundreds of Jobs in Shift Towards AI Content Moderation", Reuters, October 2024.

Exploring Alternative Approaches

Against this backdrop, strengthening the ability of law enforcement agencies to detect, monitor and respond to terrorist and violent extremist activity across complex online ecosystems has become increasingly important. Two areas offer particular promise. First, the use of OSINT methods to detect and monitor emerging threats more effectively. Second, community-driven approaches to counter terrorist and violent extremist content online, drawing on local knowledge, expertise and participation. Beyond law enforcement agencies, practitioners across civil society and small and medium-sized technology companies, also leverage OSINT methods to detect and monitor terrorist and violent extremist activity according to their mandates and capabilities, reporting such illegal content to relevant authorities. However, many of these stakeholders operate with limited access to tools and resources compared to law enforcement authorities.

Evolving OSINT Methods

OSINT – based on publicly available online sources – is an essential and widely used approach to monitoring and responding to terrorist and violent extremist exploitation of digital technologies. On the tactical level, practitioners monitor terrorist and violent extremist networks, track recruitment approaches, provide early warning around terrorist and violent extremist intent and capability, and support law enforcement investigations. When collected and stored correctly, data from open sources (particularly data from social media) serves as evidence in court against terrorism suspects.

Beyond its tactical applications, OSINT serves important strategic functions, enabling longer-term mapping, analysis and disruption of terrorist and violent extremist online ecosystems, and ongoing data collection methodologies. These inform threat assessments, prioritize threats, and determine the allocation of available resources. Such approaches must strike the correct balance between security, privacy and freedom of expression.

To assist online investigations and the preservation of digital evidence across sectors, the Rapid Response Network, known as RaReNet, and the Computer Incident Response Center for Civil Society, known as CiviCERT, have produced a free resource called the *Digital First Aid Kit*. The resource helps investigators, digital security trainers, activists and rapid responders to protect themselves against digital emergencies, including online harassment campaigns and hostile surveillance. Originally published in English, the kit is now available in Thai, Burmese, and Indonesian by EngageMedia.[26]

Approaches to investigating and responding to terrorist and violent extremist exploitation of digital technologies fall into two broad categories: manual and automated.

26 Rapid Response Network (RaReNet) and Computer Incident Response Center for Civil Society (CiviCERT), Digital First Aid Kit.

MANUAL APPROACHES

Manual investigations and monitoring – essentially human-led methods – are longstanding and widespread approaches to OSINT among law enforcement and civil society. These methods are often more cost-effective and provide greater accuracy and contextual understanding because investigators and analysts can interpret nuances and follow leads, such as following outlinks to third-party platforms. Manual investigations are particularly valuable for monitoring terrorist and violent extremist networks online. As these threat actors rapidly adapt their tactics, slang, and imagery to evade detection by automated tools, human methods remain essential. In fact, human investigators are more likely to detect any changes than fully automated tools, which sometimes struggle to detect new or otherwise previously unknown material.

Analysts use a range of tools and techniques to counter terrorist and violent extremist exploitation of the Internet. For investigations, standard approaches include Boolean queries, reverse image searches, RSS feeds and automated alerts, such as Google Alerts. Data sources include social media, messaging apps, WHOIS databases and blogs. Security tools such as virtual private networks and virtual machines are standard components of an investigator's toolkit. Effective monitoring and analysis of terrorist ecosystems requires significant expertise and resources, equipment and training, and the ability to keep pace with rapid technological change.

In spite of the benefits associated with manual investigations, they are time-consuming and lack scalability. Competition for the same funding sources among civil society organizations and think tanks in this space often results in duplication of effort and siloing of datasets. Manual investigations in the context of terrorism and violent extremism expose analysts to violent and distressing content posing significant risk to their psychosocial wellbeing. Robust organizational wellbeing policies and practices are essential to mitigate these risks. Finally, human error, such as failing to identify material that automated systems would detect, remains a persistent risk of human-led investigations.

Law enforcement agencies in some jurisdictions have sought to mitigate the scalability issue by establishing dedicated, accessible reporting mechanisms that allow members of the public, including civil society organizations, to anonymously report suspected terrorist exploitation of the Internet to the authorities. In 2018, Pakistan's National Counter-Terrorism Authority (NACTA) announced the launch of a smartphone application called Chaukas (meaning vigilant), aimed at combatting extremism and hate speech in that country. It allows users to report directly and anonymously to the relevant law enforcement agency.[27] For its part, Counter-Terrorism Policing in the United Kingdom launched a similar public app in 2021 called iREPORTit.[28] While it remains available on the Apple and Google Play stores, its uptake among the public is unclear.[29] Finally, the Royal Malaysia Police launched the Volunteer Smartphone Patrol, a public reporting tool where citizens can submit multimedia reports of criminal and terrorist-related activities with GPS coordinates directly to the National Cyber Coordination and Command Centre, providing real-time data to all sectors.[30] Human-led and community-based initiatives and tools are valuable monitoring activities that complement automated approaches, strengthening OSINT efforts through expert judgment, local knowledge, and adaptive investigative techniques.

27 Madeeha Anwar, "Pakistan Aims to Fight Hate Speech Through Smartphone App", VOA, March 2018.

28 Lizzie Dearden, "New App Launched for Reporting Terrorist Material as Extremists 'Exploit Pandemic'", The Independent, February 2021.

29 Apple, iREPORTit App.

30 Royal Malaysia Police, Volunteer Smartphone Patrol.

AUTOMATED APPROACHES

A growing number of tools and techniques now automate the detection, monitoring, and reporting or removal of terrorist and violent extremist content online. Unlike manual investigations, automation enables monitoring at scale and significantly reduces the time spent on routine and repetitive manual investigation tasks. Although expensive, once operational, these tools can reduce the cost and workload for organizations implementing automated solutions. They are particularly useful to filter relevant information from large datasets such as surfacing accounts of interest in an investigation into terrorist activity on a social media platform.

The private technology industry is leading most public sector institutions in the adoption of two predominant forms of technical tooling: hash matching and classifiers. Both tools are effective in surfacing terrorist and violent extremist content for moderation. Hash matching automatically compares multimedia content against a database of content previously identified and verified as terrorist content. These systems use hashes – a string of data generated from a piece of content that serves as its unique digital fingerprint – such as a propaganda image produced by a terrorist group. When that image is uploaded, the hash database automatically identifies – and potentially removes – a match to the existing hash on the platform.[31] Classifiers are typically based on AI models trained on a body of data, enabling them to recognize the features of specific types of content and give a score assessing the likelihood of it being terrorist content based on common identifiable characteristics. Classifiers are further discussed in the Community-driven Content Moderation section (page 15).

Two predominant forms of hashing – cryptographic and perceptual – offer different benefits and drawbacks. Cryptographic is the more secure form. It creates hashes, appearing to be a random string of data, offering nothing about the content from which it is derived. This is useful to mitigate exposure during the processing of personally identifiable information in online content. However, cryptographic hashes can only match identical copies of the hashed content. This limitation means uploads with minor edits to the material – such as the addition of text over an image – will not be automatically matched. If that is the case, cryptographic hashing may fail to detect edited versions of previously flagged content, unless exact copies of those edited versions have themselves been added to the database.

In contrast, perceptual hashing is more effective at detecting edited copies of multimedia content. Because it analyzes patterns in prominent features of an image or video, minor edits are less likely to evade detection. However, perceptual hashes are less secure than cryptographic hashes, meaning that this approach is more vulnerable to errors or privacy breaches.[32]

Both cryptographic and perceptual hashing require human input to be effective. This includes adding new content to the database or assessing the context in which hashed content appears, for example when terrorist content is posted by an academic institution or press outlet for educational or journalistic purposes. The Global Internet Forum to Counter-Terrorism's (GIFCT) flagship Hash Sharing Database uses a range of hash types and enables the surfacing and removal of terrorist content when

31 Stuart Macdonald, Ashley Mattheis and David Wells, Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online, Technical Report, January 2024.

32 Nick Locascio, "Black-Box Attacks on Perceptual Image Hashes with GANs", Medium, March 2018.

it appears on its member company platforms or when it activates its incident response protocol.[33] GIFCT reported adding "approximately 54 visually distinct items" to its hash-sharing database in response to a terrorist attack in June 2022 in Udaipur, India, in which perpetrator-produced footage of the attack circulated widely across platforms.[34] These hashes enabled the automatic removal of the content from member company platforms immediately after the attack. One of Tech Against Terrorism's flagship tools, the Terrorist Content Analytics Platform, combines manual OSINT, automated scrapers and trusted third-party flaggers to source links to terrorist content from across the Internet and then sends automated alerts about it to platforms.[35]

Many automated tools today rely on AI (artificial intelligence), which can significantly reduce law enforcement workload, particularly in managing rapidly increasing volumes of data in investigations and in automating data-heavy or time-consuming tasks. Such capabilities do not, however, come without a range of legal, ethical, and operational challenges.[36] Insights from the 2025 regional workshop in Singapore identified several use cases for AI technologies in the context of investigating terrorism and violent extremism online. AI can help automate and significantly expedite the geolocation of photos and videos; similarly, facial recognition and audio analytics are useful OSINT tools. AI can also enable investigators to identify content using sentiment analysis (based on semantics) reducing reliance on keyword searches, as well as to detect inauthentic imagery and videos.

Serious legal and ethical risks have been raised about such applications of AI, also noted during the 2025 regional workshop in Singapore, with participants particularly underscoring the slow adoption of responsible use of AI systems by law enforcement in Southeast Asia. While global resources – such as the Toolkit for Responsible AI Innovation in Law Enforcement, developed by UNICRI and INTERPOL,[37] and the CT TECH initiative guidance from the United Nations Counter-Terrorism Centre and INTERPOL – can assist national authorities, additional tailored support is needed to build operational capacities for AI use in counter-terrorism that are effective, secure, and consistent with international law and human rights obligations.[38]

33 GIFCT, "Advances in Hashing for Counterterrorism", March 2023.

34 GIFCT, "Incident Response: CI Activated in Response to Attack in Udaipur", June 2022.

35 Terrorist Content Analytics Platform, "How it works".

36 UNCCT, UNICRI, Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and Southeast Asia, 2021.

37 UNICRI, The Toolkit for Responsible AI Innovation in Law Enforcement.

38 UNCCT, INTERPOL, Guide for Establishing Law Enforcement Cooperation with Technology Companies in Countering Terrorism, 2023.

Cross-sectors Information and Intelligence-sharing

Information and intelligence sharing between sectors is a crucial component of online counter-terrorism and violent extremism prevention. Fundamentally human problems, these threats cannot be solved by technology alone. Pooling knowledge, expertise, data access and professional networks, while aligning or deconflicting objectives, can significantly increase the effectiveness of responses to these threats. Cross-sector, cross-platform, and cross-functional collaboration are key to breaking down the silos, which often hinder effective threat detection, intelligence sharing and coordinated response. This is particularly true when tackling cases connected to violent and coercive online networks, where terrorist and violent extremist content overlaps with child sexual abuse material.

From a tactical perspective, threat intelligence methods used in other sectors, such as cybersecurity and online child safety, can inform online counter-terrorism efforts, including threat detection, analysis of inauthentic or otherwise hostile networks on social media, and the investigation and analysis of website infrastructure. In this context, correlation techniques – such as linking recurring symbols, narratives, visual markers, and behavioural patterns across platforms – can infer affiliation, ideological alignment, or sympathy within online ecosystems. Cybersecurity techniques – including detecting Internet Protocol (IP) addresses behind protected infrastructure, the use of indicators of compromise, cyber threat hunting tactics, and cross-platform correlation of symbolic and relational indicators – can be applied to online counter-terrorism to differentiate meaningful signal from background noise.

Terrorist groups frequently share a similar online *modus operandi*, exhibiting patterns that closely resemble other forms of policy-violating behaviour, such as the dissemination of child sexual abuse material. These patterns include networks of linked or coordinated accounts, deliberate attempts at obfuscation and ban evasion across platforms, and the repeated use of distinctive lexicon, symbols, and media assets to signal affiliation or ideological sympathy. Despite surface-level changes in usernames, platforms, or visual presentation, these behaviours remain detectable through correlation and triangulation techniques. Initial indicators – such as a first identified account or piece of content – can be cross-referenced with lexicon analysis using natural language processing (including exact and fuzzy matching), media similarity analysis, and infrastructure-level signals such as IP addresses or client identifiers. Together, these techniques enable investigators to identify persistent behavioural and symbolic characteristics, demonstrating that even as tactics evolve, triangulation across linguistic, technical, and network-based indicators remains a highly effective method for distinguishing coordinated terrorist activity from background noise.

Cross-sector initiatives, such as The Christchurch Call Foundation,[39] GIFCT and Tech Against Terrorism,[40] have facilitated collaboration among technology companies, law enforcement agencies, civil society organizations and other stakeholders to tackle online terrorist and violent extremist content and activities. In the European Union, Europol operates a technical tool to facilitate law enforcement reporting of terrorist content to platforms called *Plateforme Européenne de Retraits de Contenus Illégaux sur Internet*. The tool facilitates cooperation between law enforcement agencies and public sector institutions in 27 European Union Member States, and contributes to effectiveness, interoperability, deconfliction, and transparency in removing illegal content from the Internet in the European Union.[41] However, these efforts require broader geographic inclusion, adaptability to the Southeast Asian context, understanding of how national law enforcement agencies operate, and consolidated information-sharing and collaboration across diverse sectors.

Several initiatives in Southeast Asia underscore the relevance of addressing online harms by providing tailored support to regional, national and local stakeholders. In Singapore, the Online Industry Safety and Security Watch Group (iSSWG) is a collaborative initiative between the Singapore Police Force and the Asia Internet Coalition, which includes companies such as Google, Grab, Facebook, Amazon, and Apple.[42] The Online iSSWG was established in 2021 to create a safer and more secure online community by promoting greater exchange of information to fight crime and counter the threat of terrorism. It aims to enhance public awareness on how to protect against scams and cyberthreats, and to encourage the sharing of expertise and best practices to strengthen the security of online platforms. In addition to these initiatives, the Online iSSWG supports the development of guidelines, frameworks and industry standards to create safer online communities.

Other cross-sector initiatives include an event in Indonesia in October 2025, organized by the Indonesian National Police Special Detachment 88 (Densus 88 Anti-Terror) and the Delegation of the European Union to Indonesia and Brunei Darussalam. The two-day collaboration brought together law

39 The Christchurch Call.

40 Tech Against Terrorism.

41 Europol, PERCI: TCO Regulation, Terrorism Working Party Presentation, September 2022.

42 Hariz Baharudin, "New Police Watchgroup to Share Counter-Terror and Cyber Crime Data with Tech Firms", The Straits Times, January 2021.

enforcement from 11 countries, including Malaysia, the Philippines, Singapore, Indonesia, and South Korea. Supported by the European Union project Enhancing Security Cooperation in and with Asia and the Indo-Pacific, known as ESIWA+, the event facilitated the exchange of threat assessments, best practices and investigative experience, including in relation to the online threat landscape.[43] With a more global approach, the CT TECH+ Project of the United Nations Counter-Terrorism Centre aims to strengthen the capacities of selected Member State law enforcement agencies in the region, including those of the Philippines and Malaysia, to respond to the use of new technologies for terrorist purposes, while upholding human rights and adopting a gender-responsive approach.[44] In particular, the project involves training on OSINT, threat assessments and policy development, as well as raising awareness on AI.

Collaborative models used in cybersecurity and online child safety provide useful examples of OSINT and information-sharing mechanisms that can inform online counter-terrorism efforts and tackle the increasing overlaps of online harm types. Information Sharing and Analysis Centers facilitate cross-sector intelligence sharing on cyberthreats and collective defence, often with a dedicated sectoral focus (e.g. healthcare, energy).[45] Shared threat intelligence platforms, such as VirusTotal, gather indicators of compromise on suspicious files, domains, IPs and URLs, to detect malware or other cybersecurity breaches, automatically sharing this information with the security community. Similarly, in the online child safety domain, the Tech Coalition is a membership organization consisting of 55 information and communications technology companies working together to share knowledge, collectively identify threats, and develop solutions to prevent child sexual exploitation and abuse online.[46] Its flagship signal-sharing tool, Lantern, helps companies to identify relevant content and activity, coordinate on child abuse-related violations on member company platforms, and report abuse to relevant authorities. The tool ingests concrete and specific signals – such as hashes, URLs, usernames, email addresses and keywords – identified as being linked to child exploitation. Other participating companies can ingest and check relevant signals against their platforms, taking appropriate action through their own enforcement systems.

Given the overlap of harm types such as terrorism, child sexual abuse material and cybercrime – particularly within violent and coercive online networks targeting youth – approaches developed in cybersecurity and online child safety have become increasingly relevant. Counter-terrorism operational capacities should be strengthened within integrated and agile institutional frameworks capable of addressing the growing convergence of online threats linked to terrorist networks and individuals, including radicalization, cybercrime, and technology-enabled exploitation such as sextortion. In this context, the establishment or strengthening of national Internet Referral Units within existing cybercrime structures can provide a multidisciplinary coordination mechanism to address illicit content across multiple crime areas, reflecting the increasingly interconnected nature of terrorism, cybercrime, and other online harms. Such arrangements can help bridge the gap between prevention and investigation, reduce institutional silos between counter-terrorism, cybercrime, and other actors playing a role in child protection and victim support, and strengthen the protection of vulnerable groups, particularly children and youth.

43 Delegation of the European Union to Indonesia and Brunei Darussalam, "Jakarta Hosts EU-Southeast Asia Counterterrorism Police Meeting Discussing Dissolution of Terrorist Organisations and Evolving Threats", October 2025.

44 UNCCT, "CT TECH+ Initiative".

45 3rd Edition of the Guide for Developing a National Cybersecurity Strategy.

46 Tech Coalition.

Community-driven Content Moderation

Currently, most content moderation on major social media platforms is the result of top-down decisions made by and within the private companies that own them. Company policies are drafted within the confines of relevant legislative and regulatory frameworks, and sometimes under political pressure, but judgment calls around borderline or subjective material, such as extremist content or hate speech, are ultimately adjudicated by the platform owners. Companies assess potentially violating content against their own policies, and navigate their own balance between freedom of expression and the mitigation of harm.

Illegal content – such as propaganda from terrorist groups that are designated in a given national context – is generally universally prohibited by platforms. However, shortcomings in internal resourcing, capacity, and expertise within some companies have meant that enforcement can be patchy, and such content originating in some regions, including Southeast Asia, has continued to proliferate. While there may be some variation at the national and regional levels, technology company policies often tend towards a uniform approach. Such an approach means differences in cultural, personal and demographic norms may not be sufficiently accommodated, and localized forms of extremism or hate may not be effectively captured or responded to.

With the growing threats posed by violent and coercive online networks, experts at the 2025 regional workshop in Singapore affirmed that existing content moderation infrastructure and efforts are largely blind to the so-called memetic forms of violence, aspirational violence consumption, and adolescent peer dynamics around violence. Platforms appear unprepared for the shift from ideological actors recruited by terrorist organizations to isolated users, often youth, inspired by and radicalized through subcultures encouraging the consumption of mass killer manifestos, attack footage, and glorification of past perpetrators. Experts warned that more efforts are required to moderate the performative element of these attacks and the potential signals posted during the preparation phase, including images of weapons and gear, obfuscated glorification of past attacks and perpetrators, and hints at operational plans. According to experts at the regional workshop, such signals are often disregarded as curiosity or morbid interest.

Top-down content moderation approaches often mean that the people who best understand local online ecosystems are largely absent from the design and implementation of moderation rules and systems. Community members often do not have a direct line of communication with big platforms and may not know how to get in contact when, for example, a local community page is taken down egregiously, or how to effectively engage a company when violent content in their local language is not being sufficiently addressed. Bilateral engagement between governments and technology companies also often leaves out the communities affected, risks over-removal, and lacks transparency in facilitating processes.

One way to empower and equip communities to contribute to content moderation processes is through organizations such as the Global Community Engagement and Resilience Fund (GCERF), a global fund dedicated to preventing violent extremism. GCERF engages local communities and strengthens their resilience against violent extremism, builds capacity, and connects local partners to national governments, foundations, and businesses. GCERF-funded projects in Indonesia have contributed to

greater community resilience, critical thinking skills, and digital literacy training in West Nusa Tenggara, Central Sulawesi, Jakarta, and West Java.[47]

Community-led approaches to content moderation can help to mitigate some local issues and concerns. Ultimate responsibility for keeping platforms safe remains with the companies that own and profit from them. Nevertheless, more bottom-up approaches enable greater user autonomy and may also help fill knowledge and capacity gaps within companies themselves. In light of this, the 2025 regional workshop in Singapore highlighted the applicability of diverse content moderation approaches in the Southeast Asian context. One suggested mechanism is to embed localized approaches within platforms and services – particularly those that operate on a centralized infrastructure, which tend to be the mainstream platforms. The "one-size-fits-none" rationale underscores the importance of localized and contextual content moderation efforts, which require approaches and settings that are scalable, customized and rooted in communities' perspectives to be effective.[48] The recent growth of trust and safety teams presents an opportunity for regional and national institutions, particularly involving civil society and expert communities, to advance customization of content moderation efforts.

Prominent examples of community-driven approaches to content moderation include features used by centralized or quasi-centralized platforms that enable community moderation, entirely decentralized and hybrid moderation models, and emerging tools and initiatives that offer future alternative approaches to top-down content moderation.

47 GCERF, Country Profile: Indonesia, October 2025.

48 Dr. Nicole Matejic, Christchurch Call Foundation, workshop presentation.

BUILT-IN FEATURES

Facilitating the ability for users to contribute to platform safety and have some autonomy over their online experience is a built-in feature of many online platforms. User reporting allows technology companies to draw on platform communities to report content that may have been missed by automated systems, such as material containing culture-specific references or slang. It is also a useful tool for end-to-end encrypted platforms or private spaces where centralized content moderation teams do not have oversight. User reporting has become a common requirement under regulatory regimes, including those in Singapore, Indonesia, and the European Union.[49]

User reporting, however, generally accounts for a small proportion of overall content moderation on larger platforms. Transparency data published for Facebook, covering April-June 2025, indicates that user reports represent 1.1% of total content removed during that period under its Dangerous Organizations policies (related to terrorism and violent extremism). For the most part, this figure has remained below 2% since transparency reporting began in late 2017.[50] Technology companies might argue that these metrics are a reflection of the effectiveness of their moderation systems. However, some users may be reluctant to report terrorist or violent extremist content when they encounter it, or may be unsure how to do so. User reporting is also unlikely to be an effective mechanism in private groups composed entirely of like-minded violent extremists, where members are highly unlikely to report their peers.

Some platforms have introduced features that allow their users to curate their online experience, including the ability to block certain types of content from their feeds. On BlueSky, a platform built on the decentralized Authenticated Transfer Protocol (AT Protocol), users have access to custom feeds and algorithmic choice, giving them control of their feeds according to personal preferences. Users can also share labels and blocklists – including for content and accounts tagged by the community as hate speech or gore – and then opt to block such content.[51] These community-led built-in features complement a company's centralized content moderation team, who oversee the enforcement of the platform's community guidelines in a top-down manner.[52]

AI IN CONTENT MODERATION: NATURAL LANGUAGE PROCESSING CLASSIFIERS

AI systems are widely embedded in content moderation efforts. Machine learning or Natural Language Processing are popular automated tools to classify terrorist and violent extremist content. These tools are typically based on AI models trained on a body of data, such as primary source text material, enabling them to recognize the features of specific types of content and give a score assessing the likelihood of it being terrorist content based on common identifiable characteristics. For the largest technology companies, content classified with a very high score of certainty is likely to be removed

49 InfoComm Media Development Authority Singapore, "IMDA's Online Safety Code Comes Into Effect", July 2023; Zulfikar Hardiansyah, "Tak Hanya Pendaftaran, Ini 4 Kewajiban Platform Digital dalam Aturan PSE Kominfo", Kompas, August 2022.

50 Meta, Community Standards Enforcement Report, Dangerous Organisation: Terrorism and Organised Hate.

51 Eli Parisier, Deepti Doshi, We Need to Protect the Protocol That Runs Bluesky, MIT Technology Review, January 2025.

52 Bluesky, Bluesky's Moderation Architecture, March 2024.

automatically, while lower scoring posts may go into a queue for review by a human moderator.

While classifiers can significantly reduce the burden of manual identification of terrorist and violent extremist content, a margin of error is inevitable, including false positives and false negatives. False positives risk the wrongful removal or data processing of legal speech and content, while false negatives may fail to identify terrorist and violent extremist activity.[53] Both these issues risk over- and under-removal of content, thereby risking the security or freedom of expression of their users. Classifiers can struggle to interpret nuance and context, particularly when content appears in language their systems do not sufficiently understand,[54] and may also carry biases inherited from their training data.

Despite several applications of AI in content moderation, efforts are scarce to use these systems effectively in non-English languages; more sustainable support is needed for the local actors aiming to fill this gap. Among the few existing examples, ShhorAI represents a noteworthy localized approach, deploying an AI-powered bot tackling hate speech across social media with a focus on Indian regional and dialect languages.[55] Built for the Indian ecosystem and designed to safeguard marginalized communities, this tool is the country's first content-moderation Software as a Service Application Programming Interface for business platforms.

CROWDSOURCING

Crowdsourcing in content moderation refers to the delegation of reviewing, filtering, and managing user-generated content to a large, distributed group of individuals rather than relying exclusively on internal staff or automated AI systems. There are several examples of crowdsourcing, with the one implemented by Wikimedia being one of the most popular. It is based on a decentralized, open-innovation model where anyone can create or edit content using MediaWiki software, turning passive readers into active contributors. This collaborative system, powered by a volunteer base, ensures quality through peer review and continuous refinement, rather than relying on a centralized and top-down structure. The Wikimedia Foundation's policy sets out clear procedures and guidelines for addressing terrorist and violent extremist content on Wikimedia projects, and thereby governs its crowdsourcing approach.[56]

Community Notes are another example of crowdsourced content moderation. In this approach, users contribute contextual labels to posts deemed false, misleading, or lacking context. The social media platform X has operated this model globally since 2021 as a crowdsourced fact-checking system. Users can apply to join the programme and, once onboarded, vote on whether notes suggested by the community will be displayed publicly on the platform. In January 2025, Meta announced in a blogpost that it would follow a similar approach, replacing its third-party fact-checking programme with a Community Notes-style approach, starting in the United States of America.[57]

53 UNCCT, UNICRI, Countering Terrorism Online with Artificial Intelligence: An overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and Southeast Asia, 2021.

54 Gabriel Nicholas, Aliya Bhatia, "The Dire Defect of 'Multilingual AI Content Moderation", Wired, May 2023.

55 ShhorAI.

56 Wikimedia Foundation, "Legal: Wikimedia Foundation Terrorist and Violent Extremist Content Procedures and Guidelines".

57 Joel Kaplan, "More Speech and Fewer Mistakes", Meta, January 2025.

Empirical evidence suggests that Community Notes can stem the spread of harmful content. A September 2025 study by the University of Washington found that posts with Community Notes attached received less engagement and were less likely to go viral than posts without. The study found that after getting a Community Note, reposts dropped 46% and likes dropped 44%, on average.[58] While Community Notes may reduce the risk of an environment conducive to radicalization, they are better suited to borderline content that does not warrant immediate removal – such as inflammatory commentary or misinformation – and are generally ill-suited to moderating content posted by members or supporters of terrorist groups.

Other research found significant shortcomings in the overall effectiveness of Community Notes on X. Focusing on posts related to the 2024 United States general election, The Washington Post found that the voting process by which Community Notes become available often exceeded 11 hours, allowing harmful content to proliferate before eventually receiving a label. It also found that a significant majority (92.6%) of the proposed notes analyzed were never published because contributors had to reach consensus before the note could be published.[59] A June 2025 study on the effectiveness of Community Notes in South Asia found a "severe linguistic imbalance" in their application, noting a strong preference for labels on English-language posts rather than those in regional languages.[60] The study authors noted the scarcity of Community Notes in South Asian languages "leaves significant information gaps, especially in fast-moving or high-stakes contexts".[61]

For Community Notes to be effective in Southeast Asia, their contributors must be diverse and representative of the region's users, a finding by Singapore-based S. Rajaratnam School of International Studies in June 2025. The researchers have argued for greater involvement of local communities by Meta in particular – given the company's recent shift to the Community Notes model – and cautioned that Community Notes should not be considered a stand-alone solution. To be effective, they must be part of a coordinated and multi-pronged effort by platforms alongside a "cross-sector network linking ethnic and faith-based organizations, journalists, third-party fact-checkers, and policymakers".[62] However, according to a report in Indicator Media, the top contributor to Community Notes on X in November 2025 was an AI Note Writer.[63]

58 Isaac Slaughter, Axel Peytavin, Johan Ugander and Martin Saveski, Community Notes Reduce Engagement with and Diffusion of False Information Online, PNAS, September 2025.

59 "Elon Musk Says X Users Fight Falsehoods. The falsehoods Are Winning", Washington Post, October 2024.

60 Center for the Study of Organised Hate, "X's Community Notes and the South Asian Misinformation Crisis", June 2025.

61 Center for the Study of Organised Hate, "X's Community Notes and the South Asian Misinformation Crisis", June 2025.

62 Lam Teng Si, Empowering Local Communities for Online Fact-Checking and Moderation, RSIS, June 2025.

63 Alexios Mantzarlis, "An AI bot is Now the Top Contributor to Community Notes on X", Indicator, November 2025.

PARTNER PROGRAMMES AND TRUSTED FLAGGERS

Many of the major technology companies[64] operate partner programmes in which specialist third-party organizations and individuals can report terrorist and violent extremist content directly to moderation teams. These programmes generally give trusted third parties certain privileges in the content reporting process, such as dedicated internal points of contact and prioritization in the content review process. While in many jurisdictions this is a voluntary process on behalf of technology companies, in the European Union trusted flaggers have been written into Article 22 of the *Digital Services Act*. Article 22 requires platforms to accommodate and prioritize reports from European Union established entities designated as trusted flaggers by European Union Member States. To ensure transparency in reporting and mitigate the risk of over-removal, these organizations are required to publish annual transparency reports on their activities.[65]

YouTube's Priority Flagger Programme – involving at least 300 government and non-government organizations globally – is a prominent example of a trusted flagger initiative. Members of the programme receive policy training from YouTube and are given access to a dedicated channel of communication with a moderation team. While trusted flagger reports are not removed automatically by YouTube, they are given prioritized status in the moderation queue.[66] YouTube does not publish the list of programme participants, though reporting in 2023 indicated that four community groups in Singapore were involved: Limitless, Samaritans of Singapore, SG Her Empowerment, and TOUCH Community Services.[67] Tech Against Terrorism's Trusted Flagger Platform is another example that enables trusted partners to submit terrorist content removal requests to technology platforms hosting verified terrorist content through a secure online platform.[68]

While these partner and trusted flagger programmes leverage vetted specialists from diverse sectors, including civil society, providing them with prioritized reporting privileges, the lack of remuneration or compensation for the reporting work has been perceived as an attempt to outsource responsibility. During the 2025 regional workshop in Singapore, concerns were raised about the difficulties of making such programmes sustainable, geographically inclusive and impactful on community guidelines and policies, particularly questioning how regional uptake can be fostered.

CONTENT MODERATION ON THE FEDIVERSE

Decentralized and federated services rely on a network of independently run servers (instances), rather than one central authority: known as the fediverse. Each instance sets its own rules, policies, and norms, resulting in a localized approach to content moderation shaped by its administrator and community moderators. In this way, users select an instance based on its localized rules and values, rather than the top-down centralized moderation approach enforced globally by a private company. Instances interact with one another unless one decides to defederate from another, such as in cases where high volumes of harmful content are appearing from a third-party server.

64 A few examples: Roblox; TikTok.

65 Article 22, "Trusted Flaggers - the Digital Services Act (DSA)"https://www.eu-digital-services-act.com/Digital_Services_Act_Article_22.html.

66 YouTube Help.

67 Joey Leong, "YouTube Priority Flagger Programme Taps Singapore Community Groups to Keep Internet Safe", Geek Culture, July 2023.

68 Tech Against Terrorism, "Trusted Flagger Platform".

Mastodon, probably the most widely known service, had around 10 million users and approximately 16,000 instances as of 2024, forming a federated ecosystem of decentralized instances, known as the 'fediverse'. Moderation efforts on Mastodon are distributed across the administrators and moderators of their respective instances, using content moderation tools offered by Mastodon's open-source software, such as blocklisting mechanisms.[69]

HYBRID APPROACHES

Some platforms operate a hybrid approach to content moderation, combining top-down rules enforced by centralized moderation teams with a decentralized community of volunteer moderators responsible for particular spaces on the platform. Reddit is an example of a hybrid approach. It maintains and enforces platform-wide policies applying to all users and communities, particularly around illegal harms such as terrorism, child sexual abuse material and hate speech. However, much of this enforcement is carried out by a network of community moderators (known as mods) who manage their own communities on the platform (known as subreddits).[70]

These volunteer mods are empowered to design and enforce their own rules within their subreddits, provided they conform with the overarching policies of the platform. To assist mods, Reddit provides an automated tool called AutoModerator, which allows moderators to automate the enforcement of their rules, such as by removing posts containing certain words or phrases or links to websites deemed inappropriate for the subreddit.[71] A similar hybrid approach is taken by the game-adjacent messaging platform, Discord, which is composed of multiple features, including the use of automated tooling for the moderation of individual servers.[72]

In summer 2025 UNICRI and UNOCT issued a call for applications as part of a project called Building Safer Online Gaming Communities, seeking to upskill community moderators, gaming community leaders, and guild chiefs based in Indonesia, Malaysia and the Philippines to more effectively counter violent extremist exploitation of online gaming spaces. Successful project applicants have contributed to research on violent extremism in gaming in Southeast Asia, receive training, and help develop a forthcoming regional guidance document on safety in online gaming spaces.[73]

MIDDLEWARE AND THIRD-PARTY FILTERING SYSTEMS

Middleware is an alternative approach to content moderation not reliant on centralized moderation teams. It is a form of third-party software that can be integrated into online platforms by users. Such tools enable users to tailor their own online experience and block or hide content they want to avoid, without relying exclusively on the features, policies and enforcement of platform owners.[74] In 2021, a middleware tool called Block Party enabled Twitter users to filter unwanted content from their feeds

69 Carlo Bono et al., "An Exploration of Decentralised Moderation on Mastodon", ACM Web Science Conference, May 2024.

70 "Cast Study: Reddit", New America.

71 "AutoModerator", Reddit.

72 "Moderation & Community Support to Manage your Server", Discord.

73 UNCCT, "Call for Young Gamers: Help Build Safer Online Communities from Violent Extremism in Southeast Asia".

74 Foundation for American Innovation, Shaping the Future of Social Media with Middleware, December 2024.

by blocking accounts engaged in spam or harassment.[75] The tool went on an indefinite hiatus in May 2023, however, following changes to Twitter's API.[76]

The market for community-driven moderation middleware is still emerging. A recent initiative in Malaysia illustrates the potential. In May 2025, Malaysian Roblox game developer MYSverse announced an additional chat filter layered over Roblox's existing filters on its flagship game, Lebuhraya. Acknowledging the limitations of "generic, international safety solutions" for "understanding local nuances", the developers said the additional filter is designed to detect and filter local Malaysian slang that might otherwise have been missed by Roblox.[77] The filter quietly obscures flagged messages without alerting the sender, making it harder to circumvent. MYSverse also uses a silent so-called shadowban system to temporarily mute disruptive players without notifying them, giving the team time to observe behaviour and make fair, well-informed moderation decisions in more complex cases. These examples are part of MYSverse's full-spectrum approach to community-led moderation, which includes proactive filtering, appeals processes, clear accountability structures and player empowerment.

Two potential routes exist by which middleware could become a viable and scalable solution in global content moderation, according to research published by the Foundation for American Innovation in December 2024. The first route would incorporate middleware into existing, centralized platforms, offering options for users to choose according to their needs. This approach would benefit from the large existing user bases of mainstream platforms, requiring no significant shifts in platform usage by Internet users. In contrast, the downside is that large platforms would retain control over the providers they use. With some exceptions,[78] large platforms have moved towards a more closed approach, reducing the chances of widespread middleware adoption by these companies in the near future.[79]

The second route for more widespread adoption of middleware favours continued growth of federated social networks. In this approach, middleware services could be built into open-source systems, such as BlueSky and Mastodon. Such an approach would not rely as heavily on the buy-in of corporate companies and would afford communities greater autonomy in self-moderating their online experience. However, this approach risks increasing the filter-bubble effect. Ultimately, the success of this approach relies on widespread migration to federated services – a market that, despite growth, still represents a small proportion of social media users globally.

75 Ian Carlos Campbell, "Block Party is Out of Beta and Ready to Block Anyone Who Likes Bad Tweets", The Verge, October 2021.

76 "Block Party's Twitter Product is on Indefinite Hiatus as of May 31", Block Party, May 2023.

77 Tan Jun Yan, "Our Commitment to Safety and Security in MYSverse", MYSverse, May 2025.

78 Naomi Mix, "Meta is Done Moderating. On Threads, Users Decide What They See", The Washington Post, July 2023.

79 Foundation for American Innovation, Shaping the Future of Social Media with Middleware, December 2024.

OPEN-SOURCE TOOLS AND INITIATIVES

Several initiatives have emerged in recent years to assist technology companies with moderating illegal content. These tools are often geared towards supporting smaller companies that may not have the same capacity or resources to identify and remove content as their mainstream counterparts. For instance, Robust Open Online Safety Tools (ROOST) is a non-profit consortium formed by Google, OpenAI, Roblox and Discord in February 2025, to create open-source platform safety tooling for platforms and developers.[80] ROOST has partnered with the Christchurch Call Foundation to respond to terrorist and violent extremist content online.[81]

In October 2025, ROOST announced a new AI-powered content classification tool developed in collaboration with OpenAI, named gpt-oss-safeguard. Publicly available and configurable according to the policies or requirements of its users, the tool can be adapted for different technical, social or legal contexts. Based on a reasoning model, the tool can explain to users why certain moderation decisions were made – a feature that is particularly useful in instances where nuance is critical.[82] Two other tools announced by ROOST in July 2025 were Coop, a content review tool with built-in integration with NCMEC, and Osprey, an open-source investigation and incident response tool.[83] Given ROOST's commitment to open-source technology and innovation, it is possible to consult all these tools.[84]

In January 2023, Tech Against Terrorism and Google's Jigsaw announced Altitude, a tool designed to help content moderators to detect, review, and remove terrorist and violent extremist content from their platforms. Specifically designed for small and medium-sized platforms, it is intended to be a cost-effective and scalable tool for community protection. To that end, Altitude is available for free on request from Tech Against Terrorism.[85]

The terrorist and violent extremist threat landscape in Southeast Asia continues to evolve, presenting significant content moderation challenges in the region. Against this backdrop, exploring alternative tools and solutions, as outlined in this report, is essential. The extent to which these content moderation tools and solutions have been adopted by technology companies of all sizes and other relevant stakeholders remains unclear. Despite this, significant efforts are required to ensure communities in Southeast Asia, and more globally, are aware of these tools and solutions and are equipped with the knowledge and skills to adapt them to local contexts, making linguistic and cultural nuances a key priority.

80 ROOST, "Launching ROOST: A Letter of Gratitude", February 2025.

81 ROOST, "Christchurch Call Partners with ROOST for Open-Source AI Tools to Tackle TVEC", February 2025.

82 ROOST , "A New Milestone for Open Source Safety Infrastructure and Transparency: GPT-OSS-Safeguard", ROOST, October 2025; OpenAI, "User Guide for GPT-OSS-Safeguard", October 2025.

83 ROOST, "ROOST Announces 'Coop' and 'Osprey': Free, Open-Source Trust and Safety Infrastructure for the AI Era", July 2025.

84 GitHub – ROOST.

85 Tech Against Terrorism, "Altitude, an Open-Source Tool Built with Google Jigsaw, is Now Under the Stewardship of Tech Against Terrorism", July 2024.

CENTRALIZED

- Top-down moderation decisions made by private companies, in line with socio-political norms and regulatory requirements
- User reporting

HYBRID

- Community moderation with top-down control and oversight
- Differing rules in different communities, in line with overarching rules

DECENTRALIZED / FEDERATED

- Localized approach to moderation, with rules set by administrators of independent instances
- Users can select instances based on their values and rules
- Instances typically linked, but can block or restrict one another

Reflections on Areas for Further Action

Discussions at the 2025 regional workshop in Singapore identified several potential focus areas for programmatic activities to build on existing OSINT methods and develop community-driven approaches to the moderation of terrorist and violent extremist content online in Southeast Asia and beyond. Based on the discussions and participants' interests, the following areas complement and enhance existing approaches and warrant consideration.

1. **Work to prevent and reduce siloed approaches for cross-harm issues.**
   Terrorism and violent extremism increasingly interact and overlap with other forms of online harm, including child safety, cybercrime, harassment, and gender-based violence. Despite the overlap, efforts to counter these threats are often conducted in isolation, and stakeholders are not sufficiently communicating with each other. Counter-terrorism approaches should better integrate collaboration, coordination, and knowledge sharing with stakeholders working on these other harm types, while integrating alternative approaches learned from these sectors. Ways to improve collaboration and coordination between actors in the counter-terrorism field, including the non-governmental sector, should be explored.

2. **Facilitate more widespread incorporation of responsible AI tooling for OSINT, particularly in the public sector and law enforcement.**
   Gaps remain in Southeast Asian law enforcement regarding knowledge and application of AI tools. Integrating AI tools significantly increases the capability of law enforcement to conduct online investigations and data analysis but increased training and deployment in this area is essential.

3. **Consider developing more red-teaming approaches for countering terrorism and violent extremism online.**
   Red teaming – the process of simulating the behaviour of an adversary to identify vulnerabilities in a target– is an effective and established practice in the field of cybersecurity, particularly for testing the security of technical architecture. While this has been applied to online counter-terrorism in some areas, such as the red teaming of chatbots and other AI tools,[86] expanding this to other areas of online counter-terrorism may help to identify gaps in moderation and other safety measures and inform further intervention opportunities. For example, simulating a vulnerable individual seeking out terrorist or violent extremist material on a given platform is likely to identify gaps in its algorithmic recommendation systems, its ability to inhibit harmful searches, and its moderation enforcement.

4. **Establish frameworks to better incorporate community voices in content moderation policies and their enforcement.**
   This should particularly involve victims, representatives of marginalized communities, or those most susceptible to the impacts of terrorism and violent extremism. Ongoing work is needed to empower these communities to contribute to the shaping of moderation decisions directly affecting them. This must be a collaborative, cross-sector effort linking ethnic and faith-based organizations, policymakers, and representatives of technology companies based in the relevant national context.

86 OpenAI, GPT-4 System Card.

5. **More resourcing should be dedicated to early intervention and support systems.**
The online threat landscape is increasingly characterized by diffuse networks of actors, rather than cohesive named organizations. This produces lone-actor perpetrators of violence, many of whom are juveniles. Given the continued availability of terrorist or violent extremist material across platforms and the growing threats posed by violent and coercive online networks, more work should be done on identifying at-risk users at the early stages of their engagement with such content. Early response mechanisms should be further developed to support these individuals, ideally involving collaboration between sectors.

6. **Encourage the integration of adaptable third-party tooling for localized content moderation solutions.**
Despite technical and commercial challenges to the widespread incorporation of middleware and third-party open-source tools in mainstream platforms, these approaches offer promising opportunities for greater autonomy and content control within online community spaces, without relying entirely on non-responsive or non-cooperative companies.


www.unicri.org