Signals of harm

An indicator-based framework to improve risk factor identification for CT and P/CVE online monitoring

CERT

POLICY PAPER

Centre for Responsible Technology

The Centre for Responsible Technology (CERT) is an autonomous research centre established by the Institute of Strategic & International Studies (ISIS) Malaysia with the foundational support of the Malaysian Communications and Multimedia Commission. CERT's mission is to provide government and key stakeholders with the knowledge and insights needed to shape technological development guided by the principle of responsibility, ensuring that such development prioritises public interest, human rights, accountability and effective regulation.

CERT's research mandate is intersectional, focusing on the relationship between technology and four primary domains: society, sustainability, safety and security. Through research, dialogue and stakeholder engagement, CERT contributes to informed policy discussions on emerging technologies and their societal implications, working at the intersection of academia, industry and government.

[POLICY PAPER]

Signals of harm

An indicator-based framework to improve risk factor identification for CT and P/CVE online monitoring

By
Anis Rifhan Rosli and Dr Haezreena Begum Abdul Hamid

Foreword by
Datuk Prof Dr Mohd Faiz Abdullah

Illustrated by
Ashikin Hussin

[AUTHORS]

Anis Rifhan Rosli is a researcher with the Special Projects team at the Institute of Strategic & International Studies (ISIS) Malaysia. She is also a member of the Malaysian chapter of the Council for Security Cooperation in the Asia-Pacific (CSCAP). Her research interests include counterterrorism, preventing and countering violent extremism, refugee management in Malaysia and the United States' foreign policy.

Dr Haezreena Begum Abdul Hamid is a criminologist and senior lecturer at the Faculty of Law, Universiti Malaya, with 24 years of legal practice. Her research focuses on human trafficking, terrorism and violent extremism, prison, incarceration, gender, human rights, sexual violence, online harms and digital victimisation. Dr Haezreena was appointed as an Expert by the International Criminal Court in counter-terrorism and human trafficking and leads the national research team reviewing AI-related reforms to the Penal Code, Criminal Procedure Code and Evidence Act.

[ACKNOWLEDGEMENTS]

We are grateful for the contributions of officials from the Royal Malaysia Police Headquarters Special Branch (E8), the Indonesia National Police and Badan Nasional Penanggulangan Terorisme, United Nations Office on Drugs and Crime Indonesia, Wahid Institute as well as numerous other organisations and individuals whose insights contributed to this research. We also extend our deepest appreciation to the participants of the expert workshops, whose contributions helped us reflect upon the key research questions.

Foreword

Among the most challenging of questions that any given society must confront is when should the state intervene before a crime has been committed. This dilemma is particularly pronounced in the context of terrorism and violent extremism, where the consequences of inaction can be catastrophic, yet premature or excessive intervention risks undermining the very rights and freedoms that democratic societies seek to protect.

This challenge takes on an added level of complexity when considering how radicalisation is no longer confined to physical spaces or through identifiable networks. Increasingly, it unfolds online through fragmented communities, algorithmically curated content and encrypted-private channels.

In this sense, individuals can encounter extremist narratives, form ideological affiliations and deepen their commitment to violent causes with nary a peep on the radar of the authorities. Policymakers are therefore confronted with a difficult balancing act: how can legitimate security concerns be addressed without eroding fundamental rights or creating disproportionate powers of surveillance?

This question cannot be answered through simplistic binaries that frame security and liberty as mutually exclusive objectives. Effective policy in this regard requires a more nuanced approach that equally weighs safety measures that are proportionate, accountable and anchored in the rule of law with the protection of individual rights without ignoring genuine threats to public security.

This paper seeks to contribute towards identifying ways to unpack this seeming binary. By examining the pathways of radicalisation and developing a framework to guide intervention, it seeks to identify when heightened scrutiny may be justified and what safeguards should accompany such actions. In doing so, it offers a pathway towards a more calibrated approach to preventing violent extremism – one that strengthens security while preserving the privacy, liberties and legal protections that underpin a democratic society.

Datuk Prof Dr Mohd Faiz Abdullah
Executive Chairman
Institute of Strategic & International Studies (ISIS) Malaysia

Abbreviations

AI	Artificial intelligence
AMLA	Anti-Money Laundering, Anti-Terrorism Financing and Proceeds of Unlawful Activities Act
BNPT	Badan Nasional Penanggulangan Terorisme (Indonesia)
CMA	Communications and Multimedia Act
CSO	Civil society organisation
CT	Counterterrorism
Daesh	Islamic State
ESI	Electronically stored information
EU	European Union
IED	Improvised explosive device
OECD	Organisation for Economic Co-operation and Development
ISA	Internal Security Act
MCMC	Malaysian Communications and Multimedia Commission
MyPCVE	Malaysian Action Plan of Preventing and Countering Violent Extremism
P/CVE	Preventing or countering violent extremism
POTA	Prevention of Terrorism Act
RMP	Royal Malaysia Police (Polis Diraja Malaysia)
SMATA	Special Measures Against Terrorism in Foreign Countries Act
SOSMA	Security Offences (Special Measures) Act
TCC	True Crime Community
TVEC	Terrorist and violent extremist content
UNODC	United Nations Office on Drugs and Crime

Executive summary

- This paper examines indicators of online extremism and proposes a structured framework to support consistent risk assessment, early intervention as well as whole-of-government and whole-of-society responses to preventing violent extremism and terrorism.

- Social media and online platforms have been exploited by extremist and terrorist groups to disseminate propaganda, spread ideological narratives and disinformation, facilitate recruitment and influence online users. They are expanding their reach and potential impacts, aided by emerging technologies.

- Gaps in Malaysia's counterterrorism and preventing or countering violent extremism efforts include difficulties in collecting legally defensible online evidence without clear thresholds, a primary regulatory and enforcement focus on clear legal violations and enforceable online behaviour as well as limited integration between regulatory and security-led online monitoring.

- The increasing complexity of online radicalisation requires a shift away from reactive enforcement towards a preventive, proportionate and evidence-based approach. One nuance is recognising that not all harmful or concerning online content necessarily meet evidential or legal standards for criminal enforcement.

- The main recommendation is to operationalise an indicator-based framework to support online monitoring. A proposed matrix classifies indicators by risk levels, matched with proportionate and recommended responses. It is complemented by a tiered risk classification system, with the option of incorporating the RAG+ and Channel systems.

- Further recommendations include reforming legal provisions, positioning restorative justice as a pre-crime intervention model, adopting a "trusted flaggers" programme, developing an accessible reporting platform with a single point of contact and strengthening cross-functional coordination.

- Policy considerations include balancing between human rights and public safety, deploying hybrid detection approaches combining artificial intelligence-assisted pre-screening with human expertise, capacity-building and clearly identifying thresholds in online content governance and criminal justice approaches.

1. Introduction

The rapid growth of online platforms has created significant challenges for regulators and enforcement agencies in identifying and responding to harmful or extremist digital content. While some online content clearly incites terrorism or promotes violent extremism, other forms may still be lawful, yet concerning.

This situation raises several important regulatory and legal questions. How can regulators and enforcement agencies distinguish between online content that incites terrorism or promotes violent extremism, and that which is legally permissible but which may still pose risks towards users and public safety? What evidence-based indicators can help authorities identify early warning signs of risk and intervene before they escalate into real-world violence? When do online behaviours cross the legal threshold for prosecution? How should authorities decide whether criminal enforcement is necessary, or whether preventive or non-criminal interventions are more appropriate?

These questions reveal a gap within Malaysia's digital safety, counterterrorism (CT) and preventing or countering violent extremism (P/CVE) framework, one which quietly shapes the national security landscape. Exposure to violent, extremist or incendiary content, whether blatant or subtle, can draw individuals towards dangerous ideas, self-radicalisation or affiliations with groups that nurture such impulses. Extremist networks understand this vulnerability well. They manipulate online spaces with calculated intent, especially by reaching the children and adolescents who form a significant portion of Malaysia's online community.[1] Only 57.5% of these young users receive parental guidance (where parents check their child's social media accounts or browser history), leaving many to navigate digital risks unprotected.[2]

Developing *evidence-based indicators and thresholds* centred upon CT and P/CVE strengthens governance, enables proactive interventions and equips civil society organisations (CSOs) and the public with tools to hold institutions accountable. In the Malaysian context, establishing common terms of reference is essential for ensuring consistent interpretation and preventing regulatory overreach, since the boundaries between harmful content, extremist expression and legally protected speech are often complex and contested. Existing legal frameworks – including provisions under the Penal Code 1936 (Act 574) relating to terrorism offences, such as Sections 130J(1)(k), 130JB and related provisions, as well as Section 233 and related provisions in the Communications and Multimedia Act (CMA) 1998 – provide mechanisms to address the dissemination of harmful or threatening content.[3]

However, the absence of clearly defined indicators for identifying early signals of extremist risk may create uncertainties for regulators and enforcement agencies. An indicator can be defined as a "behaviour that suggests an individual has likely already radicalised to violent extremism and may require more timely intervention".[4] This absence can lead to inconsistent enforcement, potential

encroachment upon legitimate expression or delayed intervention in situations where online content may contribute to radicalisation or violence. Establishing structured, evidence-based indicators would therefore support more proportionate, transparent and legally defensible regulatory responses.

Internationally, the challenge of countering violent extremism online remains unresolved, with the European Union (EU) currently leading global discussions on thresholds, indicators and intervention strategies. These efforts are reflected in its regulatory instruments, notably the Digital Services Act and the Regulation on Addressing the Dissemination of Terrorism Content Online, or Regulation (EU) 2021/784.[5] These regulations apply across all member-states and impose accountability and risk-mitigation obligations upon online platforms, including requirements to swiftly remove and report identified terrorism-related content.

Against this background, this research examines Malaysia's position within evolving global discourse, while accounting for its distinct social realities and legal complexities. It seeks to examine existing approaches and identify practical considerations for strengthening policy and regulatory responses to online violent extremism. As a disclaimer, it focuses primarily on content, rather than platform governance systems. Research employed a mixed-methods approach, combining doctrinal and non-doctrinal research with desk research and a review of academic literature, policy reports, legislation, jurisprudence and annual reports of online regulators, as well as semi-structured interviews and discussions, expert workshops and qualitative content analysis (refer also to the Appendix). This provided the foundation for identifying indicators and developing an online indicator matrix, tiered risk classification system as well as reporting and flagging mechanisms. Due to time constraints and limited engagement with some stakeholders, it was not possible to conduct interviews and discussions with all the stakeholders approached. All research participants are anonymised.

2. Malaysia's CT and P/CVE framework

"Terrorism" and "violent extremism" are pervasive terms in global security discourse and often experienced as remote concerns that gain heightened public attention during periods of media reportage on major incidents. However, there are no universally agreed-upon definitions of both terms, leading to a diversity of definitional approaches developed at the national, regional and international levels.

In Malaysia, Section 130B(2) of the Penal Code defines terrorism as "any act that is done with the intention of causing death or serious bodily injury to any person; or causing extensive destruction to a place or property; or causing serious disruption of any essential service, facility or system; or creating a public emergency."[6] The most common understanding of violent extremism, meanwhile, includes "acts or support for violence to achieve social, political or legal outcomes or in response to specific political or social grievances".[7] In contrast, there has always been little discussion of CT policy, which is primarily limited to the scholarly field.

When a public policy is introduced, it often undergoes changes and modifications over time. Hall's three "orders" of policy changes – the routine adjustments to policy settings, changes in policy instruments and shifts in overarching policy goals – can illustrate transitions in Malaysian laws regarding CT and P/CVE.[8] A third-order policy change in the CT and P/CVE framework occurred when the Internal Security Act (ISA) 1960, previously used to counter terrorism, was repealed in 2011 and replaced by the Security Offences (Special Measures) Act (SOSMA) 2012, which now serves as the procedural law for arrest, investigation and trial. Building upon this legal shift, Malaysia saw a series of first- and second-order policy changes. These included the introduction of Chapter VIA: Offences Relating to Terrorism in the Penal Code, which expanded the scope of terrorism-related offences (such as terrorist group membership, support, financing and engagement in recruitment and training for terrorist acts), and the Anti-Money Laundering, Anti-Terrorism Financing and Proceeds of Unlawful Activities Act (AMLA) 2001, which reinforced provisions on terrorism financing, freezing assets, seizure procedures and reporting requirements for financial institutions.

The factors shaping CT policymaking include interests, ideas, institutions and events.[9] Over the past 20 years, following the 11 September 2001 attacks in the United States of America and to counter the threat posed by the so-called "Islamic State" (Daesh), substantial efforts and resources have been dedicated to policymaking globally. This led to different CT responses, including in Malaysia, that required another paradigm shift in policymaking.[10] Malaysia's criminal justice system enhanced preventive detention and enforcement powers through the Prevention of Terrorism Act (POTA) 2015 and the Special Measures Against Terrorism in Foreign Countries Act (SMATA) 2015, alongside other broad measures introduced as part of the evolving CT and P/CVE framework. This includes the launch of the Malaysian Action Plan of Preventing and Countering Violent Extremism (MyPCVE).[11] The latest measure, the Online Safety Act 2025, adds a new protective layer for online

users by strengthening regulatory oversight of digital platforms, emphasising protections for children and managing online harm, by imposing obligations upon platforms that host or distribute user-generated content.[12] Although the Act requires platforms to identify, manage and mitigate risk, as well as provide reporting and user support mechanisms, there is still room for harmful content (as defined in the Act) to be shared, since it does not extend to private one-to-one messaging.

Overall, Malaysia's CT framework has evolved from an executive-based era of preventive detention to a criminal justice system characterised by legal restructuring.[13] The criminal justice system offers an institutionalised balance between the rights of the accused and society's right to security.[14] A significant turning point has also emerged in both policy and discourse, with the term "terrorism" being used less frequently, replaced by "violent extremism". These changes reflect a broader shift in focus within Malaysia's CT and P/CVE approach.

However, the comprehensive procedures for investigating and prosecuting terrorism as a criminal offence have come to be regarded as inadequate for the current digital age. This is because the nature of terrorism is evolving, and its threats are now different. Therefore, the way the government responds needs to change accordingly.

3. Definitional framework

The prevailing ambiguity in current definitions under the CT and P/CVE framework reflects their evolving and multifaceted nature, the absence of universally accepted definitions as well as variations in interpretation across jurisdictions. It is further compounded by conceptual overlaps and varying usage of terminology in existing policy and academic literature. As such, this definitional framework intends to establish operational definitions to guide analysis and ensure the consistent use of key terms.

Harmful content: Based on the Online Safety Act 2025, this is any content specified in its First Schedule. For CT and P/CVE purposes, such content is specifically defined as that which may incite violence or terrorism. Malaysia also outlines a separate definition for "priority harmful content".[15]

Internet crimes: Crimes committed on or facilitated by the use of the internet. This term is often used interchangeably with "cybercrime", which is a much broader term including any "criminal offence that has been created or made possible by the advent of technology or a traditional crime which technology's use has transformed".[16] While this term is frequently overused and includes many internet crime categories (including hacking, fraud and cyberterrorism), this paper mainly focuses on social media-enabled terrorism.

Lawful but concerning content: Also known as "lawful but awful" content or "borderline" content, it "is legal but widely considered to be socially reprehensible". Similarly to the vaguely distinguished terms "terrorism" and "violent extremism", borderline content remains an equally convoluted concept.[17]

Online content: Any digital information serving as a medium, such as comments, pictures, videos, files, posts, links, chatroom chats, blogs or messages.

Radical content: Generally, content that can shift a person's thinking and behaviour to become significantly different from how most of members of their society and community view social issues and participate politically.[18] Since radical content can be an overly broad category and may include lawful dissent, for CT and P/CVE purposes, it should be: distinguished from radical content for activism; and defined as content that creates readiness to engage in illegal and violent political action.[19]

Terrorist content: This definition should not depart from the definition of a terrorist act. For example, Regulation (EU) 2021/784 defines terrorist content as "material that solicits someone to commit or to contribute to terrorist offences, or to participate in activities of a terrorist group, incites or advocates terrorist offences, such as by glorification of terrorist acts, or provides instruction on how to conduct attacks."[20]

Violent extremist content: Violent extremist content can be defined as content that is "conceptually weaker than the term terrorism, which has an identifiable core". The Shanghai Convention includes a target requirement, while Australia and the United States include a motive requirement (such as political, religious or ideological purposes).[21] The Australian government requires three characteristics: the intention to advance a political, religious or ideological cause; supporting or facilitating serious violence; and including elements of intimidation.[22]

4. Internet-fuelled extremism and the central role of online platforms in fuelling terrorism

There are ***at least*** four routes through which online conduct evolves into real-world harm or a terrorist attack. The first is a stepwise progression through distinct stages. The second holds that not all terrorists necessarily go through distinct stages. The third holds not all terrorists go through all these stages. The fourth holds that a terrorist may not go through stages in a particular order.

Regardless of their route, individuals can engage in terrorist attacks when they have the right intent, motivation and resources, whether financial, physical and/or material (for example, access to weapons and explosives) capability; as well as research and planning (for example, hostile reconnaissance).[23] Hostile reconnaissance is understood as purposeful observation (for example, of people, vehicles, buildings, places, spaces, whether onsite or online) to inform the planning of attacks against specific targets, sometimes involving rehearsals where one or more elements of a plan are practised.[24]

Traditionally, ideology has been a significant factor in the formation of terrorist and extremist groups, by binding followers to a shared goal, providing a sense of belonging and justifying struggle. The role of the internet in radicalising individuals to violence and facilitating violent attacks through a shared ideology is not a new challenge.

One current problem is re-emerging threats of right-wing extremist ideology and religious extremist ideology (including but not limited to Islamist, Buddhist, Hindu and Christian extremisms).[25] Even white supremacist ideology has found its way to Southeast Asia in the form of "multicultural white supremacy".[26]

Anti-government extremism (including right- and left-wing variants) promoting conspiracy theories has also re-emerged in online spaces.[27] In June 2024, the Royal Malaysia Police (RMP) arrested eight people suspected of posing an imminent threat to key institutions, including the Yang di-Pertuan Agong, prime minister, dignitaries and senior police officials.[28]

Case 1: Ulu Tiram police station attack[29]

This 2024 incident in Johor was perpetrated by 21-year-old Radin Luqman Radin Imran, who was shot dead in an attack that killed two police officers and injured another. Malaysian authorities concluded that it was a "lone wolf" act, although further investigations revealed that threats could still be posed by his family members. Investigations found four air rifles, among other weapons, in the family home, with statements and phrases written on the walls by Radin Luqman, expressing hatred towards the government and security forces.

His father, Radin Imran Mohd Yassin, a former Jemaah Islamiyah member who had since pledged allegiance to Daesh, allegedly propagated its teachings to his wife and children. His wife, Rosna Jantan, was sentenced to four years' imprisonment due to her failure to provide information related to terrorism offences, and admitted that her husband had exposed the family to Daesh ideology through videos circulated by the group. The attacker's 36-year-old brother, Radin Romyullah Radin Imran, had downloaded video clips of Daesh from social media, fantasised about attacking the Merdeka Day parade in a van packed with explosives, and identified police officers and military personnel as targets. The court sentenced Radin Imran and Radin Romyullah to 30 years' imprisonment each, after they pleaded guilty to terrorism-related charges.

Case 2: State High School 72 school bombing in Indonesia[30]

On 7 November 2025, a 17-year-old student bombed his high school with improvised explosive devices (IEDs), having drawn inspiration from far-right extremists and school-shooters, injuring 96 individuals. Post-attack discussions described the case as a combination of memetic violence and True Crime Community (TCC) subculture, inspired to a minor extent by white supremacist attacks in the West. This is a key example of how youths can be influenced to commit attacks triggered by online content without adhering to a definite ideology.

Another challenge is that virtual connections and platforms hosting violent extremist content and connecting individual supporters of extreme ideologies are getting vast, diverse and, most of the time, hidden, thus ensuring that the internet continues being the main medium for terrorist and extremist networks.[31] In these digital spaces, online users seek validation and search for identity, where like-minded individuals and groups engage with content that promotes violent extremist beliefs, sentiment, propaganda and disinformation.

Social media and communication platforms such as X, YouTube, Facebook, Telegram, Reddit, TikTok, 4chan and Gab are commonly used by extremist and terrorist groups.[32] The anonymity affordances of these online platforms and their transcendence of geographical boundaries are conducive for communication and operational coordination. In recent years, terrorists have kept up with platform innovations by amplifying impact and glorifying attacks via instant broadcasting (for example, through videos, infographics or social media livestreams). The first case of terrorist

livestreaming occurred during the 2019 attack on Christchurch mosques in New Zealand (the first terrorism conviction in its history), exposing problems in regulating and moderating abhorrent content on social media.

Another development is the likelihood of "lone wolf" attackers, who are either self-radicalised or self-initiated terrorists.[33] Such individuals are not radicalised by a terrorist group directly. Ideology and personal grievances may inspire them, but they otherwise work alone, while actively seeking out extremist material online to justify their worldviews. A limited exposure to diverse perspectives can contribute to the risk of self-radicalisation upon encountering or engaging with extremist content.

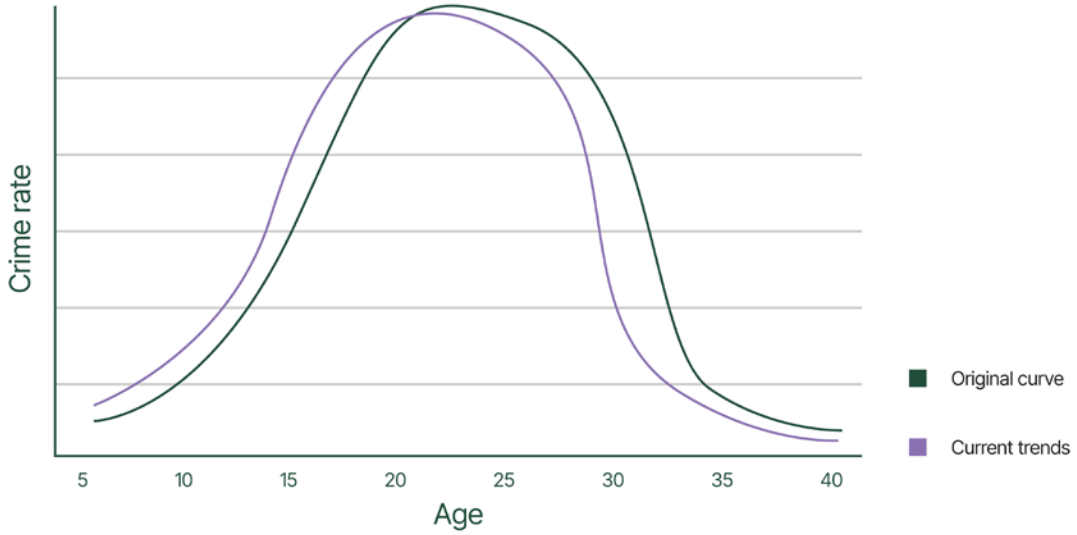
Case 3: Christchurch Mosque terrorist attack[34]

> On 15 March 2019, 28-year-old Australian terrorist, Brenton Tarrant, a white supremacist, livestreamed his attack on Facebook Live using a headcam. Before the attack, he was an active user of fringe online forums, who posted a 74-page "manifesto" online espousing various neo-fascist aims. The "manifesto" stated that the attack had been planned two years earlier, and that Christchurch had been scoped out three months in advance. He started the livestream during his drive to the first mosque attack, while listening to a song idolising Radovan Karadzic (jailed for genocide against Bosnian Muslims). He opened fire at the Al Noor Mosque and Linwood Islamic Centre using rifles, targeting worshippers during Friday prayers. 51 people were killed and 40 people sustained gunshot injuries. Tarrant was convicted of terrorism and is now serving a life imprisonment sentence without parole.

Another concerning global trend is the rising number of young people engaging in violent extremist ideologies.[35] Due to the prevalence of violent extremist content online, combined with personal vulnerabilities (whether emotional, physical or social), youths are more easily exploited by violent extremists and terrorist groups.[36] For example, social isolation can lead teenagers to seek social outlets offering a sense of belonging. In March 2026, a large police operation placed at least 97 Indonesian youths (the youngest aged 10) under police monitoring after they engaged with content glorifying physical and memetic violence, having been inspired by TCC discussions.[37] Their electronically stored information (ESI) were found mostly on messaging applications such as Telegram and social media networks like X.[38] Here, minimal regulatory constraints produced ideal environments for terrorist networks and extremist groups to propagate their ideologies.

The average time taken for youth radicalisation used to be 18 months in 2005, contracting to 13 months in 2016, and as of 2026, just a few weeks.[39] Based on current trends, the age-crime curve, whose peak should be in the late-teens or early-20s, appears to be shifting leftwards relative to the normal.[40]

Fig. 1: The age-crime curve[41]



Case 4: Singapore youth idolising white supremacist ideology[42]

In 2025, an 18-year-old Singaporean youth, Nick Lee Xing, was influenced by far-right extremism via an online game, where he role-played as a terrorist killing Muslims in a mosque. He idolised the Australian terrorist, Brenton Tarrant (see Case 3) and wanted to attack Muslims at the mosque. Lee was found to have developed a hostile view towards Muslims in early 2023 after encountering Islamophobic and far-right violent extremist content on social media. Based on police investigations, he got a tattoo and T-shirts with custom prints of logos associated with neo-Nazi, far-right and white supremacist groups. On 10 February 2025, an order of detention was issued by the Internal Security Department, under Singapore's ISA.

Case 5: Malaysia teen "lone cub" attack[43]

On 11 January 2016, the police arrested a 16-year-old teenager in Sungai Petani, Kedah, claiming Daesh affiliation, after he tried to kidnap a Chinese sales assistant at a shopping complex, holding her hostage with a knife. He admitted that he was ordered to conduct the attack by Malaysian Daesh followers, and that his actions sought to publicise Daesh's existence to the Malaysian authorities and public. The teenager was also radicalised to view non-Muslims as "kafir harbi" (those who can be justifiably killed). It was stated that the teenager was brainwashed by email and social media, and was eager to prove that he was capable of such an act.

Whether online users encounter, consume or engage with content directly through a terrorist group, or indirectly via algorithm, the risk of radicalisation remains constant unless appropriate measures are implemented. Causality may be hard to prove, however, because of the complex factors underlying individual actions. Regardless of whether individuals are radicalised through online connections or if those drawn to online forums are already radicalised, such content requires action. After all, it remains a risk factor in pathways facilitating engagement with and potential exploitation of others, therefore creating broader ripple effects within online networks.[44] Accordingly, even without asserting a direct causal relationship between internet use and radicalisation or violent extremism, this paper maintains that exposure to specific categories of online content warrants policy attention.[45]

5. Overview of current gaps and operational limitations

5.1 Difficulties in collecting legally defensible online evidence without clear thresholds

Investigating online crimes related to CT and P/CVE requires more than just the ability to look for information. While online content moderators may only need training to identify specific types of harmful content before taking further action, law enforcement agencies must collect and preserve evidence in a manner consistent with its admissibility in court. For each case, guilt must be proven beyond a reasonable doubt and to a specified standard of evidence or proof. Therefore, ESI must be relevant and authentic (and not hearsay) to qualify as admissible evidence.

Before a police officer can begin preparing an investigation paper, there must be sufficient and reasonable grounds of suspicion. Online ESI is "collected from the Internet, which has been properly collected, preserved and for which a foundation has been laid for its admittance and/or consideration into a legal forum", while digital ESI is "collected from computers or electronic media which has also met all the legal requirements for admittance and/or consideration into a legal forum".[46] While digital ESI can be seized, online ESI is collected through rigorous online monitoring and live acquisition.[47] Online data collection exists only temporarily and could disappear or be deleted unless captured and stored in a certain manner (usually as screenshots of ESI at a particular date, time and internet protocol address, collected from a website, social networking site, instant messaging application or chat box).[48] Besides documentation alone, the frequency and corroborative weight of ESI from the same user also matters (as evidenced by a series of screenshots).[49]

In applying this approach to suspected terrorism-related offences online, police officers may need to invest significant time compiling and timestamping screenshots to establish clear legal violations, while making sure that there is no tampering or manipulation prior to data collection via hashing.[50] However, compiling screenshots is no longer a valid online documentation method in and of itself: the determination of protocols, collection methodologies and locations also matter.

Fig. 2: Online ESI collection steps[51]

Without clear thresholds, it is challenging to determine the corroborative weight of individual indicators, or combinations of indicators, arising from pre-investigations and during online monitoring. The authorities may face difficulties in determining whether or not the available evidence is sufficient to initiate a formal investigation and appropriate action. Without a clearly defined threshold, a pre-investigation may be subject to inconsistent assessment, risks subjective or biased conclusions and draws public scrutiny to matters of transparency, fairness and procedural consistency. However, delays caused by uncertainties in decision-making should not be overlooked in matters of national security and public safety.

5.2 A primary regulatory and enforcement focus on clear legal violations and enforceable online behaviour

Under the criminalisation approach, a higher standard of evidence is required to bring cases to court.[52] Since the current framework is designed to respond primarily to legal breaches, institutional attention is often directed towards behaviours that satisfy the legal threshold for enforcement. For example, gradual behavioural shifts, such as constantly liking, commenting upon and sharing harmful posts on social media, may receive less attention than postings of extremist content.

Additionally, many forms of "lawful but harmful" content exist within regulatory grey areas. Preventive intervention mechanisms remain underdeveloped when potential indicators of radicalisation are diffuse, context-dependent and inadequately captured by online monitoring units, despite their cumulative influence on radicalising vulnerable users.

Given the significance of subjective interpretation or vague guidelines, existing regulatory and enforcement systems may also face constraints in differentiating between freedom of expression (which may include engagement in controversial discourse) and genuine indicators of mobilisation towards violent acts. For example, an individual may consume or share potentially harmful content, but does not necessarily demonstrate an intention to commit unlawful conduct at a particular time.

This situation can be managed effectively with clear thresholds and appropriate interventions for individuals at particular risk levels. There is also growing recognition of the need for more risk-based frameworks in CT and P/CVE approaches to complement current legal enforcement mechanisms.

5.3 Limited integration between regulatory and security-led online monitoring

Currently, there is no single, comprehensive source of guidelines or step-by-step process for law enforcement agencies and regulatory bodies to refer to when addressing harmful online content related to CT and P/CVE. Standardised guidelines or step-by-step process flows will enhance integration between the law enforcement agencies and regulatory bodies overseeing online monitoring.

However, there is currently limited integration and cooperation between regulatory oversight mechanisms and security-led online monitoring.[53] Regulatory bodies and law enforcement agencies operate in separate ministries and are given separate mandates, possibly resulting in fragmented approaches to addressing online radicalisation risks. Differences in operational priorities and evidentiary thresholds may also hinder information-sharing. Consider how regulatory approaches tend to focus on platform compliance and the responsible use of technology, while law enforcement agencies prioritise threat-identification, public safety and national security. For example, regulatory bodies may prioritise making online digital spaces and social media safer for public use, while law enforcement agencies prioritise the prevention of real-life attacks, preferring that online footprints be maintained for ease of online monitoring.

6. Research findings on extremist and terrorist exploitation of digital platforms

6.1 Algorithmic amplification of grievance content

Social media algorithms rank content based on users' likes, comments, shares and other interactions, customising recommendations and amplifying content with higher engagement to maintain users through feedback loops.[54] This can occur on social media and video-sharing platforms (such as Facebook, YouTube, X, Gab and Reddit), where basic interest or random searches can lead to more consumption of related online feeds via "filter bubbles".[55] For example, when a potential recruit views violent extremist content online, algorithms can create an "echo chamber" where only content that reinforce the presentation of extremist narratives, propaganda and ideologies are fed to this user in the future.

In other words, the algorithms have been weaponised to gradually steer vulnerable individuals towards extremist and terrorist ideologies through emotional resonance. This parallels the exploitation of emotional vulnerabilities and personal crises of online users (such as feelings of alienation and/or specific grievances due to personal struggles) by terrorist groups to recruit followers. After the 2019 Christchurch terror attack, New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron formed The Christchurch Call, with its commitment to "review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist and violent extremist content".[56]

6.2 Migration from mainstream platforms to encrypted spaces

Mainstream and closed platforms as well as internet forums are often key enablers and gateways for promulgating violent ideologies. Usually, interactions between a terrorist group and potential targets begin on mainstream platforms (for example, WhatsApp, Telegram and Discord). Having fostered personal interactions, kinship and trust, terrorist groups often migrate conversations to encrypted spaces to reduce visibility to law enforcement and impede investigation.[57]

These closed spaces foster a sense of safety, anonymity, active participation and personal/community interaction, where active recruitment and indoctrination usually start undetected.[58] Encrypted messaging applications often use end-to-end encryption to deter third-party entities and even developers from accessing users' chat boxes and chatrooms.[59] Applying rational choice theory to this context, terrorist and extremist groups prefer this method due to its low risk (risk), ease of technological facilitation (effort) and capacity for building online reputations among subcultural groups by claiming responsibility for certain terrorist actions (reward).[60]

6.3 Humour, satire and memes normalising violence

To avoid violating platform policies, another communication strategy executed by terrorist and extremist groups is coding harmful content through humour, satire and memes, which take advantage of vague definitions of borderline content as well as borderline terrorist and violent extremist content (TVEC) content.[61] TVEC, a term often used by platforms to merge both terrorist and violent extremist content under one umbrella category, is especially useful for reporting purposes.

"Saint Hoax", the popular meme creator, defines a meme as "a piece of media that is repurposed to deliver a cultural, social or political expression, mainly through humour. It can capture insight in a way that is in complete zeitgeist".[62] Often taking the form of "image memes", it spreads quickly via the internet, including online platforms. Just as memes can accelerate the popularity of celebrities, politicians and forms of entertainment, the same dynamic applies to extremist and terrorist content.

Due to their virality and accessibility, memes as well as humorous and satirical content can be weaponised to propagate toxic narratives, symbols and messages by forming a subcultural and memetic language, doxing or promoting groupthink (closely related to extreme ideologues engaged in trolling or producing propaganda, harmful stereotypes, derogatory content or abusive and offensive materials).[63] Most importantly, the original intention can be hidden and thus openly shared in mainstream domains.[64]

6.4 Exploitation of gaming platforms and gaming communities

Violent extremist groups increasingly use online gaming chat rooms to identify, recruit and radicalise young people via in-game chat rooms and game-agnostic (not game-specific) social platforms. The use of such media or offline messaging (store-and-forward messaging) to normalise narratives associated with violent extremism target youths in particular.[65] Immersive gaming technologies (such as virtual reality and games taking place on the metaverse) that incorporate violent behaviour may possibly lead vulnerable youth to normalise violence, although no clear correlation has been established.[66] Regardless, games amplifying certain political and ideological stances may blur the lines between creativity in virtual spaces and lived social deviance.

Gaming platforms and communities such as Roblox, Gorebox, Discord, Fortnite and Minecraft are being exploited to radicalise young people, including Malaysians. In a recent case, six individuals, including teenagers, were detained, having been believed to have pledged allegiance to the leader of a Daesh-linked extremist organisation.[67] Online misinformation and disinformation may further amplify violent extremist narratives. Left unchecked, this may erode trust in public institutions and democratic society, manifesting as offline violence and terrorist acts.

6.5 The use of the deep web and dark web by terrorist groups

To circumvent regulatory constraints, terrorist groups use the dark web to plan and coordinate attacks, disseminate propaganda and recruit new followers.[68] In general, the deep web can be understood as password-protected databases and intranets hidden from the usual surface web. Meanwhile, the dark web includes sites that can only be accessed via specialised browsers, which most users will never interact with on a daily basis.[69] The anonymity and convenience offered by the dark web as well as various underground markets and forums create an enabling environment for planning attacks. Dark web marketplaces are where cybercrime-as-a-service thrives, in which lone wolf attackers and terrorist groups can easily access auxiliary services (weapon procurement, encryption and secure communication channels) hidden from online monitoring by the authorities.[70]

Terrorist groups utilise the dark web to acquire independent domains (.onion) and publish their URLs on the surface web, deep web, encrypted communication platforms such as Telegram or specialist dark web cybercrime forums.[71] These URLs can be disseminated among followers to share training and instructional materials in the form of online courses, instructional videos, e-books, e-manuals and webinars.[72] However, law enforcement agencies and content moderators face difficulties extracting information offline for investigation.

6.6 Behavioural escalation markers

Observable shifts in online behaviour that may indicate a gradual progression or escalation of engagement, radicalisation or hostile reconnaissance can be detected through online monitoring.[73] Such markers may include discussions on operational security, evasions of law enforcement officials, praise for successful extremist attacks by glorifying attackers or their modus operandi, joining online groups focused on promoting extremism and violence (such as chat groups discussing IED manufacturing), virtual simulations of attacks or the use of unusual and powerful encryption tools in online messaging applications.[74] The migration of individuals from mainstream digital spaces and social media to encrypted online spaces is another relevant marker.[75]

Recognising these key behavioural and emotional changes in beliefs and attitudes, online behaviours and social patterns after certain types of online consumption will greatly assist in official assessment and intervention. Escalatory behavioural markers may illustrate a transition away from passive content consumption towards active participation within the terrorist and extremist ecosystem. However, many online behaviours remain ambiguous and may overlap with other purposes. Therefore, they require further context-dependent assessment. For example, law enforcement and regulatory bodies need to know how to assess persons expressing extremist views or stronger political stances online, by determining whether these constitute healthy debate or the adoption of extreme views.

6.7 Distortions of religious narratives

This is not a new challenge in Malaysia. It consists of the manipulation, selective interpretation or misrepresentation of religious teachings to legitimise or justify extremist ideologies and worldviews, while demonstrating a group's religious credentials.[76] Ideology exploits religious concepts, symbols and terminologies to frame violence as morally and religiously justified.[77] Based on reviews and resolutions by United Nations bodies – ranging from the General Assembly and Security Council to its Global Counter-terrorism Coordination Compact entities and the International Criminal Police Organisation – there has been "a rise in terrorism on the basis of xenophobia, racism and other forms of intolerance or in the name of religion or belief".[78]

Take for example the Al-Qaeda-linked Jemaah Islamiyah, which had its roots in Malaysia and the region. Established in 1993, it was the product of an earlier radical movement, Darul Islam, which ideologically espoused an "Islamic state".79 Operating regionally through loosely linked cells, among Jemaah Islamiyah's deadliest attacks was the 2002 Bali bombing that killed more than 200 people. As of now, 16 of its leaders have announced the group's disbandment by video statement.[80] However, remnants of the network remain active, and post-disbandment does not prevent future attacks under its continued ideological influence.[81]

Another concern is the use of artificial intelligence (AI) tools by terrorist and extremist networks, which may leverage charismatic online influencers or religious authorities to increase the persuasive reach of their narratives. Exploitation can include media spawning, automated multilingual translation, fully synthetic propaganda, the repurposing of old propaganda using generative AI tools and personalised propaganda.[82] Additionally, geopolitical developments local grievances and sociopolitical tensions may be leveraged to facilitate the distortion of religious narratives, framing them as evidence of an existential threat to particular religious communities.

7. Policy recommendations

Having explored online extremist content as well as the potential role of the internet and digital spaces in facilitating violent extremism and terrorism, the need for a strategic approach to prevent and investigate such activities and criminal conduct is imperative. Several recommendations are described below, which should be complemented by appropriate legal mandates for law enforcement agencies and regulatory bodies to conduct investigations and interventions.

7.1 Operationalising an indicator-based framework for CT and P/CVE online monitoring to improve consistency in identifying risk factors

An indicator-based framework for CT and P/CVE online monitoring will enhance the ability of regulatory bodies and law enforcement agencies to recognise non-obvious patterns indicating gradual exposure to extremist narratives.[83] This framework will improve the identification of contextual warning signs by linking shifts in behavioural markers to online engagement, social influence and risk levels of radicalisation. It supports a more holistic assessment approach, in which the combination and identification of multiple indicators enhances pre-crime assessment procedures, rather than relying on isolated incidents or expressions.

This framework also strengthens the ability to detect commonly used explicit, implicit and concealed or coded extremist messaging in digital environments. It facilitates more informed decision-making by investigators and reviewers by grounding assessment in documented indicators. In cases requiring early intervention, this framework supports the identification of vulnerable individuals who are exposed to extremist and terrorist ecosystems.

Risk levels are divided into four tiers (passive, medium, high and critical), which are then matched with relevant online indicators and recommended responses (see also Recommendation 7.2).

Table 1: Suggested online indicator matrix

Examples of online indicators	Risk level	Recommended response
TIER 1		
Repeated consumption of martyrdom content	Passive	Education or engagement through counter-narratives
Discriminating against other individuals of groups of people		
Expressing views that affect social cohesion		
TIER 2		
Being drawn towards conspiracy theories that reinforce extremist worldviews	Medium	Monitoring
Repeated violent threats combined with extremist fixation		
Accessing extremist or terrorist websites, forums and publications		
Expressing an interest in travelling to a conflict zone		
Expressing persistent intolerance towards groups of people perceived as "Other", including through exclusionary, discriminatory or demeaning rhetoric and general calls for violence against such groups		
TIER 3		
Sharing violent propaganda	High	Intervention
Searching for secure communication techniques and methods to communicate with terrorist groups		
Engagement with encrypted recruitment channels		
Obsession with and admiration for massacres, extreme or mass violence (for example, through online searches, membership in TCC groups)		
Being in contact with groups or individuals known to support violent extremist ideologies		
Direct attempts to establish personal contact with known extremist actors		
Obsession with virtual simulations of attacks		
TIER 4		
Hostile reconnaissance	Critical	Criminal enforcement
Operational attack planning		
Taking part in any proscribed violent extremist group's activities (for example, fundraising, sharing materials, recruitment)		
Producing or disseminating terrorist material		
Obtaining and purchasing blueprints for weapons (for example, 3D-printed guns)		

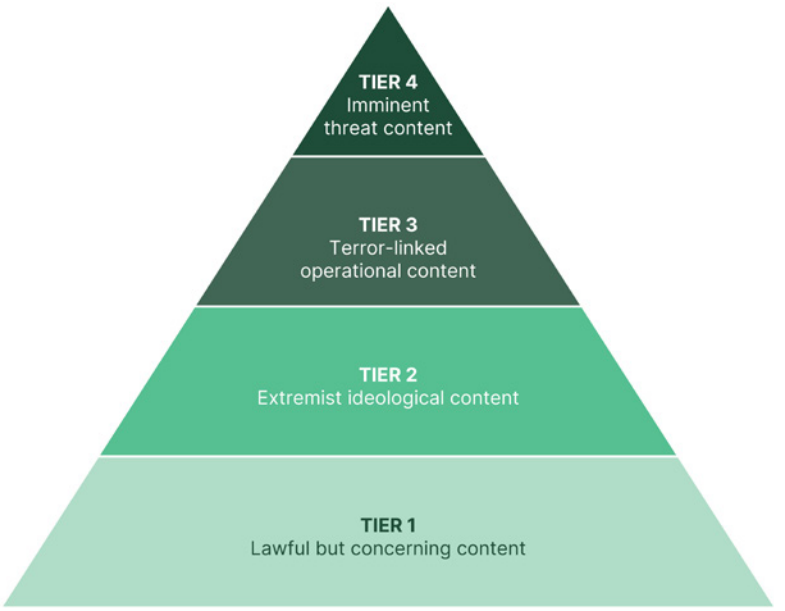
7.2 Introducing a tiered risk classification system to guide proportionate responses

Individual experiences of engaging and interacting with harmful content vary in salient ways.[84] The policy and practice of pre-crime intervention and mitigation should identify and address variations in online risk. A tiered risk classification system enables stakeholders to identify and consider tailored interventions, depending on an individual's risk level, type of online behaviour, spaces occupied, membership in communities and levels of engagement.[85] It also helps determine an individual's level of culpability for online behaviour.

Based on Figs 3–5, the lowest tier, Tier 1 (passive risk) involves engagement with lawful but concerning content; Tier 2 (medium risk) involves extremist ideological content; Tier 3 (high risk) involves terror-linked operational content; and the highest, Tier 4 (critical risk), involves imminent threats. These tiers are operational classifications for risk assessment and intervention purposes and the terms used *should not be conflated* with those introduced in the definitional framework.

Tier 1 indicates that there is no clear evidence that individuals are susceptible to violent extremism or terrorism, and further observation or waiting for recurring behaviour may be warranted. Tier 2 indicates that individuals may be vulnerable to translating online behaviours into real-world harm, and thus investigators need to continue online monitoring more closely and regularly. Tier 3 indicates that individuals are at high risk of criminal action, where law enforcement agencies should be alerted for intervention. Tier 4 indicates that individuals may commit violence at any opportunity, thus requiring criminal enforcement (for example, intelligence gathering, questioning, search and seizure, immediate arrest, detention, charging and prosecution).

Fig. 3: Recommended tiered risk classification system

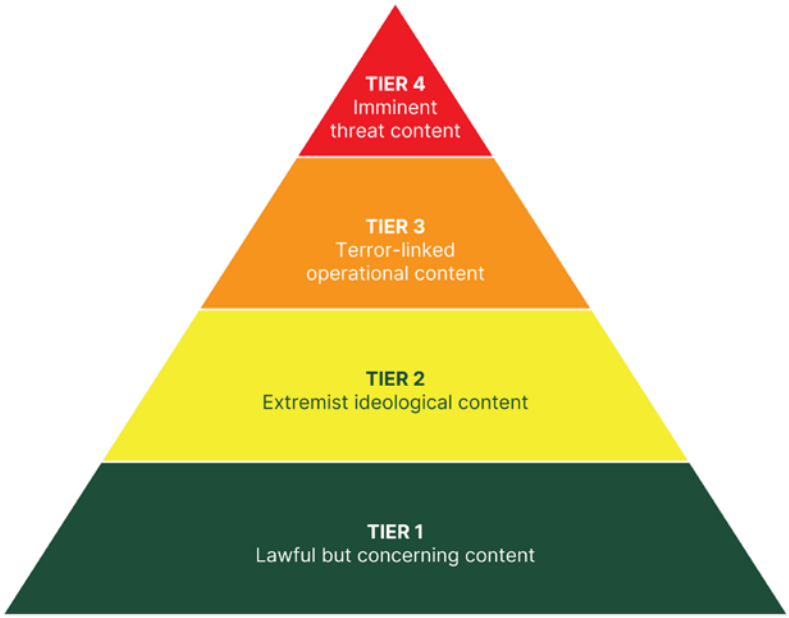


It will be helpful to incorporate and merge the colour codes of the RAG+ system (also known as the traffic light tiered assessment). Measurable information is classified by colour: red, amber, green and an additional colour, usually orange. In fields like medicine, education and natural disaster management, this system has reduced uncertainty in assessment and improved decision-making.[86]

To further enhance the classification system, an approach modelled after Channel (a British multi-agency programme designed to provide early intervention for individuals most at risk of being drawn into terrorism) can help structure a holistic approach by incorporating elements of identification and intervention.[87]

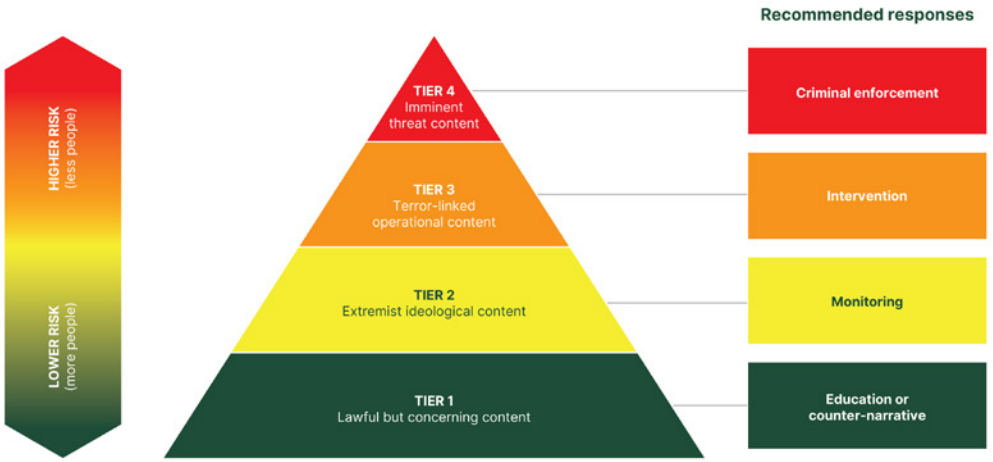
Once the higher thresholds (belief in an individual's general involvement or tendency towards terrorism-related activity) are crossed, authorities should proceed with intervention and/or criminal enforcement. However, if the thresholds are not met, online users are categorised into lower tiers, where law enforcement agencies need to implement forward-looking measures to predict and prevent possible future offences.

Fig. 4: Proposed incorporation of the RAG+ system into a tiered risk classification system



Based on Fig. 5, Tier 1, for example, appropriate pre-crime intervention to counter extremist online postings include improving public awareness about the risks of extremism, misinformation and disinformation as well as the responsible use of technology. This can be done through technology literacy education, talks on CT and P/CVE and even participation by the authorities in workshops or public programmes explaining the recruitment methods and tactics employed by terrorist and extremist groups. Former extremists and terrorists can also be invited to provide first-hand insights into the dangers and risks of radicalisation.[88] Those at risk of radicalisation should benefit greatly from the different perspectives offered here.

Fig. 5: Proposed incorporation of the Channel approach into a RAG+ tiered risk classification system



7.3 Reforming legal provisions to clarify thresholds for online prosecution and enforcement

There is a need to reform Section 130JB of the Penal Code, which concerns the possession of items associated with terrorist groups or acts. This section states that "whoever has possession, custody or control of, or provides, displays, distributes, or sells, any items associated with any terrorist group or the commission of a terrorist act shall be punished with imprisonment for a term not exceeding seven years, or with fine, and shall also be liable to forfeiture of any such item."

Due to its overly broad and vague definition of possession, there is a risk of unintended consequences, such as selective enforcement targeting political dissenters or activists, violations of freedom of expression and the right to information as well as unjust prosecution. Further, unlike other offences in Chapter VI (which includes terms like "knowing", "intentionally" and/or "having reason to believe"), Section 130JB lacks a clear provision for a *mens rea* (intention) element.

Given the strict liability nature of this section, the extent to which courts may consider a person's intentions or contextual purposes for possessing materials in question is limited. For example, in *Siti Noor Aishah Atam v. Public Prosecutor (2018)*, the Kuala Lumpur High Court found that under Section 130JB(1)(a), she was guilty of having in her possession 12 books allegedly related to terrorist and terror activities. However, she contended that she was using these books for educational purposes.[89] Similar cases, in which the accused were charged with the possession of terrorism-linked items, include *Public Prosecutor v. Muhammad Sani Mahdi Sahar (2016), Public Prosecutor v. Muhammad Hakimin Azman (2017), Public Prosecutor v. Azizi Abdullah (2017) and Mohamad Nasuha bin Abdul Razak v. Public Prosecutor (2019)*.

Therefore, a modification to this section should incorporate a *mens rea* requirement as a proactive disclosure mechanism. Statutory carve-outs for *bona fide* activities (acting sincerely without intent to deceive or break the law) will effectively resolve existing ambiguities and align the provision with its original legislative intent.

7.4 Exploring a restorative justice pilot programme as a pre-crime online intervention model

Unlike punitive CT approaches involving criminalisation, there is a need to consider whether or not such measures are necessary for users who encounter or engage with harmful online extremist content, but remain in the early phases of radicalisation. With the introduction of MyPCVE and rapid proliferation of online extremist material, mechanisms operating *outside* the formal criminal justice system targeting low-level offenders and non-operational actors are needed. Singapore, Denmark and Indonesia have each implemented rehabilitation and diversion-oriented approaches to address radicalisation risks.

Recognising restorative justice as a pre-crime intervention model for online extremism sends a clear message to the public that the system is not designed to punish, but rather, to prevent further action by individuals exhibiting online curiosity that may indicate early-stage vulnerability to extremist content and narratives, those who may translate online behaviours into violent extremism and plan to commit terrorist offences as well as those suspected of future involvement in terrorism.[90] The nature of restorative justice mechanisms is determined by what is considered necessary for preventing involvement in terrorism and violent extremism.

Besides the intervention of law enforcement agencies through this approach, appropriate resources to address other issues leading youths to engage with harmful content should be made available to support early intervention. A whole-of-society approach is pertinent for implementing restorative justice mechanisms across a variety of cases involving vulnerable individuals.[91] This means including anyone (for example, parents, teachers, community leaders and trusted peers) deemed important for helping return potential offenders to positive behaviours before they engage further with harmful content.[92] For example, a young adult who

engages with extremist content to distract themselves from dealing with a family breakdown may require a family counsellor's services, but another one dealing with trauma may need a mental health therapy session from a psychiatrist or psychologist.

7.5 Appointing "trusted flaggers" and developing an accessible reporting platform as a single point of contact

CSOs, experts actively conducting CT- and P/CVE-related initiatives and research as well as advocates for the responsible use of technology should be engaged to bridge knowledge gaps, promote information-sharing and enable the proactive identification of harmful content.[93] This step will not only increase the capacity of human moderators and reviewers to detect harmful content, but also improve the transparency of the monitoring system and response times for removal requests.[94] For example, through Indonesia's "Trusted Flagger" programme, volunteers (mostly from the CSOs like Wahid Institute, ICT Watch and Anti-Defamation of Society) are given the authority to flag content violating a website's terms of service or community guidelines.[95]

In addition, an online reporting platform should be made available for CSOs and the public to easily and safely report concerning behaviours or incidents to a single point of contact within the government. Simplification will enhance annual CT and P/CVE risk assessments and illustrate the exact magnitude of challenges faced. "Trusted flaggers" can also use this platform to escalate concerns around any content deemed capable of radicalising or triggering online users to commit violence.[96]

To fully understand the scope of threats, it is recommended that the communications and online regulatory actor further explores avenues to improve and document data collection (for example, establishing working groups with law enforcement agencies and coordinating with the private sector, such as platform regulators, social media companies and CSOs).

7.6 Strengthening cross-functional coordination for timely responses and standardised intervention pathways

Key stakeholders should advance information-sharing across agencies, including through stronger partnerships between the regulatory body and law enforcement agencies' online monitoring and CT units, by focusing on identifying and managing harmful content. Malaysian Communications and Multimedia Commission (MCMC) data indicate that major platforms removed approximately 92% of the 697,061 posts flagged as harmful between January 2024 and November 2025, although some 58,104 posts were still accessible at the time of reporting due to the use of automated and AI tools which hindered detection efforts.[97] Posts that are still accessible may not yet meet the threshold for criminal enforcement,

but through cross-functional coordination, online users identified engaging with these posts can be referred to standardised intervention pathways.

In the field of CT and P/CVE, law enforcement and regulatory bodies tend to think of transparent communication as reducing public support. However, transparency over decision-making rationales may actually *increase* policy support and public trust, especially in controversial cases.[98] Cross-functional coordination enhances transparency and accountability between different agencies, builds public trust by explaining rationales behind efforts and educates the public in making informed decisions about social media and online platform usage.

Strong coordination between a regulatory body and online platforms, in relation to transparency over TVEC, will inform law enforcement agencies', relevant authorities' and policymakers' respective threat assessments. For example, the MCMC can enhance its annual online content supervision report by including the aggregated complaints received (or content flagged) on TVEC and offer updates on the online safety landscape in consideration of CT and P/CVE.[99] However, significant data collection is needed. Based on Organisation for Economic Co-operation and Development (OECD) data, only 17 of the top 50 most popular online content-sharing services issued transparency reports with information on TVEC in 2024, and of the top 50 services used to spread TVEC online, only six did so.[100]

By creating a more cohesive and coordinated approach, stakeholders from both law enforcement agencies and regulatory bodies can collectively enhance their abilities to navigate and respond effectively to the dynamic landscape of internet-enabled violent extremism and terrorism.

8. Policy considerations

8.1 Safeguarding freedom of expression and human rights

CT and P/CVE invite considerable discussion and debate about the balance between safeguarding national security and freedom of expression, between public safety and individual liberties.[101] Consequently, intervention frameworks should recognise the heterogeneity of possible ideological expressions involving the same type of online content. Each reported case should be treated contextually, rather than applying a one-size-fits-all rule.[102] Newly introduced policies and frameworks should strike the right balance by avoiding under- or over-reactions that may lead to low public trust and criticism of official efficiency.

Since most investigations entail the collection and processing of personal information, measures taken must account for human rights-compliant implementation. Robust safeguards, consistent with international human rights norms and standards, must be applied, while opportunities for representation as well as access to appeal mechanisms and judicial review should be provided.[103]

To find this balance, stakeholders should understand that simply engaging with violent extremist content online *does not necessarily* lead to a risk of violence or amount to terrorist activity. Therefore, when addressing issues within the general category of online safety, stakeholders should opt for a harm prevention approach, rather than focusing on countering violent extremism.[104]

8.2 Integrating local cultural and contextual sensitivities into risk assessment and profiling practices

Cultural context matters.[105] Even seemingly similar cases of online radicalisation or extremism require different responses in different countries, even those within the same region. This is because "significant institutional differences are likely to develop in each country, because a different set of initial conditions produces a different set of actions and events which have a cumulative effect and set institutional development on a different path".[106]

Due to the sheer volume of posts, comments, videos and images uploaded every second, a manual review by human moderators and reviewers alone is impractical. With the growing scale of digital platforms and user-generated content, a hybrid moderation model – combining multilingual AI tools and a diverse moderation team – can combine the speed and scalability of AI with the intrinsic skills of human reviewers (for example, having contextual awareness of local cultures, social norms, religious sensitivities, ethical judgement).[107] The empathetic nature of human reviewers can never be underestimated. It is impossible to implement efficient CT

and P/CVE strategies by solely utilising AI, even with advancements and emerging technologies. For example, the impact of doxing in relation to the 2024 KK Mart incident requires a contextual understanding of the transition from online to offline transgression, in relation to online posting and online vigilantism.[108]

Human reviewers are especially needed to detect extremist content, especially in non-English languages. In Malaysia, this includes content not just in the national language (Bahasa Malaysia) or major languages such as Mandarin and Tamil, but also numerous indigenous languages (numbering over 100) and those of other communities. Additionally, dialects vary significantly across states and communities due to local histories and geographies. Taking all this into account, human capabilities may be limited, but the limitations faced by AI in relation to non-English extremist and terrorist content is greater. Left solely to an automated system, harmful elements will never be captured and continue circulating on social media.[109]

8.3 Increasing the competency and capacity of online monitoring units and content moderators

To handle internet crimes, stakeholders in online monitoring units and content moderators must possess foundational CT and P/CVE proficiency, cybersecurity expertise and technical acumen. Specific training is not just advisable but essential to empower them with the knowledge and skills required to conduct terrorism-related investigations, identify relevant online content and consider human rights dimensions.

While not every team member needs to become a CT and P/CVE expert, a foundational knowledge base is indispensable. This ensures that employees across various agencies can communicate effectively and collaborate seamlessly. For example, a government content moderator should possess enough knowledge to engage and exchange meaningful insights with CT personnel. Utilising advanced technological skills in terrorism-related investigations also helps speed-up investigations (therefore requiring the continuous improvement of AI systems). Some factors to consider include the ease of committing acts of terror or other crimes using technology, ongoing difficulties in tracking cybercriminals and suspected offenders online as well as challenges in collecting digital evidence.

This does not mean that law enforcement and regulatory bodies are not progressing. In fact, the RMP's online monitoring unit has been proactively investigating cases, analysing digital and online evidence as well as compiling evidence related to online radicalisation. On the other hand, the MCMC, as the sole regulator under the recent Online Safety Act, leads investigations, enforcement and other actions against platforms that fail to protect users. While the criminal justice system intervenes at the level of legal culpability, content governance prioritises content moderation and regulation.

8.4 Thresholds in online content governance and criminal justice

Having extreme beliefs does not automatically lead one to violence, and the majority of those who support and subscribe to content or act as bystanders online do not commit offline violence.[111] It is important that regulatory bodies and law enforcement agencies acknowledge that not all harmful or extremist content should be criminalised. Therefore, rather than focusing on criminalisation, this paper has aimed to improve the regulatory framework by introducing an indicator-based framework and tiered risk classification system.

Most individuals falling in Tiers 1, 2 and 3 should not necessarily be arrested or prosecuted, since law enforcement agencies undertake criminal enforcement actions only at Tier 4, when imminent threats are detected. This can only be done efficiently through the coordination of all stakeholders. For example, government content moderators should report only imminent threats to the RMP once risks are detected and classified as Tier 4.

Further analysis on online CT and P/CVE criminal enforcement is beyond the scope of this paper. However, it is important to determine, for example, how police officers can issue charge sheets to suspected offenders whose online behaviours fall in Tier 4 (that is, initiating criminal proceedings). A causal link to offline harm is required, subject to clear evidentiary standards based on the Evidence Act 1950, without pre-empting outcomes. This involves identifying and clarifying appropriate legal and evidentiary thresholds for determining when online behaviour demonstrates a sufficiently substantiated causality with offline harms.

9. Conclusion

The internet has become a pervasive and invasive part of everyday life, extending from private life into the public realm. This policy paper has assessed the wide range of behaviours that can be exhibited in online spaces, and considers whether or not a person holding radical or extremist views necessarily translates them into violent offline behaviours and terrorist acts. While social media platforms employ measures such as content moderation, user-reporting tools to enforce community standards as well as automated systems and de-platforming, terrorist groups and extremist networks adapt their strategies to undermine these initiatives.

The emerging picture of radicalisation has become complex and requires preventive measures, but current enforcement approaches mostly remain reactive rather than preventive. Therefore, Malaysia needs a proportionate framework that distinguishes between exposure, vulnerability and actual mobilisation towards violence. The risks posed by online extremist content to society should be tackled through hybrid solutions incorporating online and offline policies, strategies and early intervention, while accounting for proportionality and rights protections to prevent over-criminalisation. Multi-stakeholder collaboration and cross-sectoral initiatives are needed to improve consistency in regulating online content, by bridging government, civil society and industry actors. Finally, content regulation should no longer operate as an opaque process, but instead be grounded in transparent risk indicators, evidence-based and consistent assessment criteria, proportionate interventions calibrated to clearly defined risk tiers and accessible reporting platforms.

Appendix: expert workshop participants and contributors

Participants included representatives from various ministries and agencies, CSOs and industry, for instance. The views expressed during the workshops should not be interpreted as official positions or endorsements.

Government ministries and agencies

1. Attorney General's Chambers
2. CyberSecurity Malaysia
3. Institute for Youth Research Malaysia
4. Malaysia Defence Intelligence Organisation
5. Malaysian Islamic Development Department
6. Malaysia Prisons Department
7. Defence Ministry
8. Digital Ministry
9. Education Ministry
10. National AI Office
11. National Security Council
12. Prime Minister's Department (Research Division)
13. RMP Headquarters (Crime Prevention and Community Safety Department)
14. RMP Headquarters (Special Branch)
15. Southeast Asia Regional Centre for Counter Terrorism
16. Social Welfare Department

Civil society, academic and professional bodies

1. IMAN Research
2. Initiative to Promote Tolerance and Prevent Violence (INITIATE.MY)
3. ISIS Malaysia
4. Khazanah Research Institute
5. Malaysian Bar (cyber and privacy committee)
6. Malaysian Youth Diplomacy (MyDiplomacy)
7. Malaysian Institute of Defence and Security (MiDAS)
8. Universiti Kebangsaan Malaysia (National University of Malaysia)
9. Universiti Malaya

Industry / digital platforms

1. ByteDance
2. Meta Platforms, Inc.

Key discussions

1. Types of harmful and extremist content available online that may influence radicalisation and promote violent extremism;

2. Potential risks and harms posed by such content to vulnerable populations, particularly children and youth, through radicalisation, recruitment and engagement of extremist behaviours;

3. Evolution in the radicalisation process and factors contributing to the escalation of online extremism into real-world violence;

4. Structured early-warning indicators to identify exposure to harmful and radicalising content and support timely preventive interventions; and

5. Feasibility and desirability of an assessment framework to detect content.

Endnotes

[1] Institute for Economics and Peace. (2025). Lone Wolf and Youth Terrorism: Evolving Patterns of Terrorism & Radicalisation in Western Democracies. Vision of Humanity. https://www.visionofhumanity.org/resources/?type=research; and Tan, C. T. (2024, August 23). Young men most vulnerable to extremist content, says group. Free Malaysia Today. https://www.freemalaysiatoday.com/category/nation/2024/08/23/young-men-most-vulnerable-to-extremist-content-says-group

[2] UNICEF. (2020). Executive Summary: Malaysia Edition: Our Lives Online: Use of social media by children and adolescents in East Asia—opportunities, risks and harms. https://www.unicef.org/malaysia/reports/our-lives-online

[3] Penal Code (Act 574). (Malaysia). https://lom.agc.gov.my/act-detail.php?act=574&lang=BI

[4] Department of Homeland Security (USA). (2021, August 28). FAQ Sheet: What are risk factors and Indicators? https://www.dhs.gov/publication/risk-factors-and-targeted-violence-and-terrorism-prevention

[5] Anonymous officers, UNODC, personal communication, February 25, 2026; EU. (2021, April 29). Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0784&utm_source=perplexity; and European Commission. (n.d.). The Digital Services Act. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act

[6] Penal Code (Act 574).

[7] Government of Australia. (2025). A Safer Australia: Australia's Counter-Terrorism and Violent Extremism Strategy 2025. https://www.nationalsecurity.gov.au/what-australia-is-doing/a-national-approach/australias-counter-terrorism-strategies

[8] Hall, P. A. (1993). Policy Paradigms, Social Learning, and the State: The Case of Economic Policymaking in Britain. Comparative Politics, 25(3), 275–96. https://doi.org/10.2307/422246

[9] Renard, T. (2021). Counter-Terrorism as a Public Policy: Theoretical Insights and Broader Reflections on the State of Counter-Terrorism Research. Perspectives on Terrorism, 15(4), 2–10.

[10] Mat Rus, M. (2025, July). The (Over)Criminalisation of Terrorism Offences in Malaysia and Indonesia (ICCT Policy Brief). The International Centre for Counter-Terrorism. https://icct.nl/publication/overcriminalisation-terrorism-offences-malaysia-and-indonesia

[11] Bernama. (2024, September 30). Govt Launches MyPCVE Action Plan To Counter Terrorism Intelligence, Extreme Activities. https://bernama.com/en/news.php?id=2346318

[12] Bernama. (2026, February 4). Online Safety Act 2025 Strengthens Regulatory Oversight of Digital Platforms. https://www.bernama.com/en/news.php/?id=2519824

[13] Mat Rus, M. (2022). Malaysia's counter-terrorism policy: Shifting from the executive-based to the criminal justice approach? UUM Journal of Legal Studies, 13(1), 409–29. https://doi.org/10.32890/uumjls2022.13.1.16

[14] McCauley, C. (2009). War versus criminal justice in response to terrorism: the losing logic of torture. In W. G. K. Stritzke et al. (eds.), Terrorism and Torture: An Interdisciplinary Perspective (pp. 63–85). Cambridge University Press.

[15] Online Safety Act 2025. https://lom.agc.gov.my/act-detail.php?act=866&lang=BI

[16] Shipley, T. G., & Bowker, A. (2014). Investigating Internet Crimes: An Introduction to Solving Crimes in Cyberspace. Syngress.

[17] Deedman, J. (2023, May 3). The Internet Consortium for Online Safety: How collaborative tech, not legislation, could prevent harmful content proliferation. Global Network on Extremism & Technology. https://gnet-research.org/2023/05/03/the-internet-consortium-for-online-safety-how-collaborative-tech-not-legislation-could-prevent-harmful-content-proliferation/

[18] UNDP. (2022). Preventing and Countering Violent Extremism (PCVE) in Malaysia: Handbook for Civil Society Organisations (CSOs). https://www.undp.org/malaysia/publications/preventing-and-countering-violent-extremism-pcve-malaysia-handbook-civil-society-organisations-csos

[19] Moskalenko, S., & McCauley, C. (2009). Measuring Political Mobilization: The Distinction Between Activism and Radicalism. Terrorism and Political Violence, 21(2), 239–60. https://doi.org/10.1080/09546550902765508

[20] European Commission. (2021, April). Factsheet on Terrorist Content Online. https://home-affairs.ec.europa.eu/document/download/506de61d-c53f-489f-a9e3-e16e78793e81_en?filename=202104_terrorist-content-online_en.pdf

[21] Vaughan, K. (2022). The Interoperability of Terrorism Definitions: GIFCT Legal Frameworks Working Group. Global Internet Forum to Counter Terrorism. https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-LF-TVEC-1.1.pdf

[22] Attorney-General's Department (Australia). (n.d.). Violent Extremist Material offences (VEM). https://www.ag.gov.au/national-security/publications/violent-extremist-material-offences

[23] Action Counter Terrorism. (n.d.). Spotting the signs of terrorism. https://act.campaign.gov.uk/spotting-the-signs-of-terrorism/

[24] Australian National Security. (n.d.). Hostile reconnaissance. https://www.nationalsecurity.gov.au/protect-your-business/crowded-places/crowded-places-guidelines/hostile-reconnaissance

[25] Protect UK. (2023, February 28). The Threat from Extreme Right-Wing Terrorism. https://www.protectuk.police.uk/threat-risk/threat-analysis/threat-extreme-right-wing-terrorism; and Singam, K. V. (2025). Extreme Right's Influence and Impact on Southeast Asia. In K. Y. Ong & D. D. Cheong (eds.), Violent Extremism: What Was Learnt, Where We Are, and What's Next (pp. 111–16). World Scientific.

[26] Ramakrishna, K. (2025). Jemaah Islamiyah, ISIS and Beyond. World Scientific.

[27] Byman, D. & McCabe, R. (2025, September 23). Left-Wing Terrorism Is on the Rise. The Atlantic. https://www.theatlantic.com/politics/archive/2025/09/charlie-kirk-left-wing-terrorism/684323/; National Coordinator for Counterterrorism and Security (Ministry of Justice and Security, the Netherlands). (2022, November 7). NCTV's Terrorist threat assessment: threat in and to the Netherlands has become more multifaceted and diffuse. https://english.nctv.nl/latest/news/2022/11/07/nctvs-terrorist-threat-assessment-threat-in-and-to-the-netherlands-has-become-more-multifaceted-and-diffuse; and Ware, J. (2026). Western Far-Right Terrorism in 2025 and Beyond. Counter Terrorist Trends and Analyses, 18(1), 127–33;

[28] Bernama. (2024, June 24). Eight arrested for suspected Islamic State links; investigations reveal threats to Malaysia's king, PM Anwar. Channel News Asia. https://www.channelnewsasia.com/asia/eight-arrested-islamic-state-links-malaysia-king-pm-anwar-threats-radicalism-extremism-4432371

[29] Asyraf, F. (2026, April 28). Ulu Tiram attacker's brother fantasised about Merdeka Day attack, court hears. Free Malaysia Today. https://www.freemalaysiatoday.com/category/nation/2026/04/27/ulu-tiram-attackers-brother-fantasised-about-merdeka-day-parade-attack-court-hears

[30] Ghaisany, S. A. I., & Basha, S. (2026, January 29). SMAN 72 School Bombing in Indonesia. Jamestown. https://jamestown.org/sman-72-school-bombing-in-indonesia/

[31] Dass, R. (2025, February 11). The Continued Threat of Online Radicalisation in Malaysia. The Diplomat. https://thediplomat.com/2025/02/the-continued-threat-of-online-radicalization-in-malaysia/

[32] Rothut, S. et al. (2026). Radicalization and the internet: 25 years of (online) radicalization research. The Communication Review (advance online publication). https://doi.org/10.1080/10714421.2026.2658935

[33] Anonymous officer, BNPT, personal communication, February 26, 2026.

[34] BBC. (2020, August 27). Christchurch mosque attack: Brenton Tarrant sentenced to life without parole. https://www.bbc.com/news/world-asia-53919624; Martin, L., & Smee, B. (2019, March 15). What do we know about the Christchurch attack suspect? The Guardian. https://www.theguardian.com/world/2019/mar/15/rightwing-extremist-wrote-manifesto-before-livestreaming-christchurch-shooting; and Royal Commission of Inquiry into the terrorist attack on Christchurch Mosques on 15 March (New Zealand). (n.d.). Summary of the report of the Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019. https://christchurchattack.royalcommission.nz/publications/report-summary

[35] Anonymous officer, BNPT, personal communication, February 26, 2026.

[36] Channel News Asia. (2023, February 21). Two self-radicalised Singaporean boys given ISA orders; 15-year-old youngest to be detained. https://www.channelnewsasia.com/singapore/internal-security-act-isa-self-radicalised-singaporeans-students-detained-islamic-state-al-qaeda-roblox-3292691; and Royal Canadian Mounted Police. (n.d.). Five-Eyes Insights – Young people and violent extremism: a call for collective action. https://rcmp.ca/en/corporate-information/publications-and-manuals/five-eyes-insights-young-people-and-violent-extremism-call-collective-action

[37] Anonymous officer, BNPT, personal communication, February 26, 2026; and Teresia, A., Widianto, S., & Potkin, F. (2026, March 10). White Supremacist content grips teens plotting attacks in Southeast Asia. Reuters. https://www.reuters.com/business/media-telecom/white-supremacist-content-grips-teens-plotting-attacks-southeast-asia-2026-03-10/

[38] Anonymous officers, RMP, Special Branch Counterterrorism Division (E8), personal communication, April 21, 2026; and Shipley & Bowker (2014).

[39] Institute for Economics and Peace. (2026, March 30). Global Terrorism Index 2026. https://www.economicsandpeace.org/reports/; and Sabur, R. (2026, March 19). Children radicalised "within weeks" by online extremism. The Telegraph. https://www.telegraph.co.uk/news/2026/03/19/children-radicalised-weeks-online-extremism-social-media/

[40] Anonymous officers, Indonesia National Police (Densus 88), personal communication, February 25, 2026; and Crimpsy. (2025, April 6). The Age-Crime Curve in Criminology: Understanding Patterns of Criminal Behaviour. https://www.crimpsy.com/the-age-crime-curve-in-criminology/#Understanding_the_Age-Crime_Curve

[41] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026.

[42] Tan, C. (2025, February 10). Singapore teen influenced by far-right held by ISD, wanted to attack Muslims in a mosque. The Straits Times. https://www.straitstimes.com/singapore/sporean-student-18-with-far-right-beliefs-pretended-to-kill-muslims-in-online-games-held-by-isd

[43] Tan. G. C., et al. (2016, January 13). Teen stages IS "lone cub" attack. The Star. https://www.thestar.com.my/news/nation/2016/01/13/teen-stages-is-lone-cub-attack/

[44] Aryaeinejad, K., & Scherer, T. L. (2024). The Role of the Internet and Social Media on Radicalisation: What Research Sponsored by the National Institute of Justice Tells US. Department of Justice (US). https://www.ojp.gov/ncjrs/virtual-library/abstracts/role-internet-and-social-media-radicalization-what-research

[45] Whittaker, J. (2022). Online radicalisation: What we know. European Commission, Radicalisation Awareness Network (Publications Office of the EU). https://home-affairs.ec.europa.eu/system/files/2023-11/RAN-online-radicalisation_en.pdf

[46] Shipley and Bowker (2014), 73.

[47] Anonymous officers, RMP (Special Branch), personal communication, April 21, 2026.

[48] Ibid.

[49] Ibid.

[50] Vanini, C., et al. (2025). Understanding strategies and challenges of timestamp tampering for improved digital forensic event reconstruction. In Proceedings of the Digital Forensics Doctoral Symposium (pp. 1–8). https://doi.org/10.1145/3712716.3712727

[51] Anonymous officers, RMP (Special Branch), personal communication, April 21, 2026.

[52] Shipley & Bowker (2014).

[53] Bernama. (2026, March 7). MCMC to obtain details from police on teen terrorism case. https://www.bernama.com/en/news.php?id=2531354

[54] Awasthi, S. (2025, February 13). From clicks to chaos: how social media algorithms amplify extremism. Observer Research Foundation. https://www.orfonline.org/expert-speak/from-clicks-to-chaos-how-social-media-algorithms-amplify-extremism

[55] Whittaker, J. et al. (2021). Recommender systems and the amplification of extremist content. Internet Policy Review, 10(2). https://doi.org/10.14763/2021.2.1565

[56] The Christchurch Call. (n.d.). Understanding Algorithms and Developing Interventions. https://www.christchurchcall.org/understanding-algorithms-and-developing-interventions/

[57] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026; and Anonymous officers, RMP (Special Branch), personal communication, April 21, 2026.

[58] Ogele, E. P. (2025). Unveiling the shadows: Exploring the roles of the dark web and encrypted messaging apps in facilitating online terrorist networks. Journal of Strategic and Global Studies, 8(2), art. 3. https://doi.org/10.7454/jsgs.v8i2.1175

[59] Ogele (2025).

[60] Choi, K. S., Lee, C. S., & Cadigan, R. (2018). Spreading propaganda in cyberspace: Comparing cyber-resource usage of Al Qaeda and ISIS. International Journal of Cybersecurity Intelligence & Cybercrime, 1(1), 21–39. https://doi.org/10.52306/01010418ZDCD5438

[61] Department of Internal Affairs (New Zealand). (n.d.). Countering violent extremism: What is terrorist and violent extremist content? https://www.dia.govt.nz/Countering-Violent-Extremism-What-is-terrorist-and-violent-extremist-content; and Ware, J. (2023). The third generation of online radicalization (Program on extremism, George Washington University). https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/2023-06/third-generation-final.pdf

[62] Benveniste, A. (2022, February 14). The Meaning and History of Memes. New York Times. https://www.nytimes.com/2022/01/26/crosswords/what-is-a-meme.html

[63] Anonymous officers, Wahid Institute, personal communication, February 26, 2026; and Martinez Pandiani, D. S., Tjong, E. K. S., & Ceolin, D. (2025). "Toxic" memes: A survey of computational perspectives on the detection and explanation of meme toxicities. Online Social Networks and Media, 47, 100317. https://doi.org/10.1016/j.osnem.2025.100317

[64] Schwarzenegger, C., & Wagner, A. (2018). Can it be hate if it is fun? Discursive ensembles of hatred and laughter in extreme right satire on Facebook. Studies in Communication | Media, 7(4), 473–98. https://doi.org/10.5771/2192-4007-2018-4-473

[65] Devaraj, S. (2026, March 8). Rehabilitation professionals warn of the dangers of online platforms in radicalising youth. The Straits Times. https://www.straitstimes.com/singapore/politics/rehabilitation-professionals-warn-of-dangers-of-online-platforms-in-radicalising-youths

[66] Directorate-General for Migration and Home Affairs (European Commission). (2024, August 27). RAN small-scale expert meeting – New technologies and P/CVE: threats and opportunities for practitioners. https://home-affairs.ec.europa.eu/whats-new/publications/ran-small-scale-expert-meeting-new-technologies-and-pcve-threats-and-opportunities-practitioners-26_en

[67] Camoens, A. (2026, March 6). Police warn parents: Extremists are targeting young gamers on Roblox. New Straits Times. https://www.nst.com.my/news/nation/2026/03/1391175/police-warn-parents-extremists-are-targeting-young-gamers-roblox-watch?source=widget

[68] UN. (2019). Secretary-General Calls Cyberterrorism Using Social Media, Dark Web, 'New Frontier' in Security Council Ministerial Debate. https://press.un.org/en/2019/sgsm19768.doc.htm

[69] Kaspersky. (n.d.). What Is the Dark Web? https://www.kaspersky.com/resource-center/threats/deep-web

[70] UN Office of Counter-Terrorism. (2024). Beneath the Surface: Terrorist and violent extremist use of the dark web and cybercrime-as-a-service for cyber-attacks. https://unicri.org/beneath-surface-terrorist-and-violent-extremist-use-dark-web-and-cybercrime-service-june-2024

[71] Murphy, H. (2021, Sept 16). Telegram emerges as new dark web for cyber criminals. Financial Times. https://www.ft.com/content/cc3e3854-5f76-4422-a970-9010c3bc732b?syn-25a6b1a6=1

[72] Ogele (2025).

[73] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026.

[74] UNDP. (2022). Preventing and Countering Violent Extremism (PCVE) in Malaysia: Handbook for Civil Society Organisations (CSOs). https://www.undp.org/malaysia/publications/preventing-and-countering-violent-extremism-pcve-malaysia-handbook-civil-society-organisations-csos

[75] Anonymous officers, RMP (Special Branch), personal communication, April 21, 2026.

[76] New Zealand Security Intelligence Services. (n.d.). New Zealand's Security Threat Environment 2025. https://www.nzsis.govt.nz/our-work/new-zealands-security-threat-environment/security-threat-environment-2025/view/3354

[77] Yeo, K., & Rijal, A. S. (2024, November 18). Youth and Adolescent Online Radicalisation: Critical Cases from Singapore. Global Network on Extremism & Technology. https://gnet-research.org/2024/11/18/youth-adolescent-and-online-radicalisation-in-singapore/

[78] UN General Assembly. (2022, August 3). Report of the Secretary-General: Terrorist attacks on the basis of xenophobia, racism and other forms of intolerance, or in the name of religion or belief. (Document A/77/266). https://docs.un.org/en/A/77/266

[79] Reuters. (2024, July 4) Southeast Asia armed group Jemaah Islamiyah to disband: Report. Al Jazeera. https://www.aljazeera.com/news/2024/7/4/southeast-asia-armed-group-jemaah-islamiyah-to-disband-report; and SG101. (2025, October 6). Terrorism and Self Radicalisation. https://www.sg101.gov.sg/defence-and-security/current-threats/terrorism-and-self-radicalisation/

[80] Reuters (2024).

[81] Zenn, J. (2024, July 31). Radin Luqman: Police Stabbing in Malaysia Raises Questions About Future of Recently Disbanded Jemaah Islamiyah. Jamestown. https://jamestown.org/brief/radin-luqman-police-stabbing-in-malaysia-raises-questions-about-future-of-recently-disbanded-jemaah-islamiyah/

[82] Nelu, C. (2024, June 10). Exploitation of Generative AI by Terrorist Groups. The International Centre for Counter-Terrorism. https://icct.nl/publication/exploitation-generative-ai-terrorist-groups; Tech Against Terrorism. (2023, November). Early terrorist experimentation with generative artificial intelligence services. https://techagainstterrorism.org/hubfs/CT%20Tech%20Trends%20Report%20-%20Early%20Terrorist%20Experimentation%20with%20Generative%20Artificial%20Intelligence%20Services.pdf

[83] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026.

[84] Aryaeinejad & Scherer (2024).

[85] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026.

[86] Engeset, R. V., et al. (2022). Colours and maps for communicating natural hazards to users with and without colour vision deficiency. International Journal of Disaster Risk Reduction, 76, 103034. https://doi.org/10.1016/j.ijdrr.2022.103034; Örnek, F., et al. (2024). Traffic Light Tiered Assessment: An example from STEM. Online Science Education Journal, 9(2), 30-43. https://izlik.org/JA39PU58DT; and Unity Health. (2020, December 16). Study find traffic light system helps reduce clinical uncertainty and improve doctors' treatment decision. https://unityhealth.to/2020/12/study-finds-traffic-light-system-helps-reduce-clinical-uncertainty-and-improve-doctors-treatment-decisions/

[87] Child Law Service. (n.d.). Radicalisation in schools and the Prevent Duty. https://childlawadvice.org.uk/information-pages/radicalisation-in-schools-and-the-prevent-duty/#elementor-toc__heading-anchor-7

[88] Anonymous officers, Indonesia National Police, personal communication, February 25, 2026; and Anonymous officer, RMP (Special Branch), personal communication, February 17, 2026.

[89] Gunaratnam, S. (2017, April 26). Siti Noor Aishah jailed 5 years for possessing terror-related materials. New Straits Times. https://www.nst.com.my/news/crime-courts/2017/04/233975/siti-noor-aishah-jailed-5-years-possessing-terror-related-materials

[90] Maring, L. (2026). Why restorative justice is not punishment (and why the distinction matters). Journal of Social Philosophy. https://doi.org/10.1111/josp.12615

[91] Anonymous officers, Wahid Institute, personal communication, February 26, 2026.

[92] Aziz, F. (2024, May 27). Review child justice system, urge experts. The Star. https://www.thestar.com.my/news/nation/2024/05/27/review-child-justice-system-urge-experts; and Devaraj (2026).

[93] Anonymous officers, Wahid Institute, personal communication, February 26, 2026.

[94] Ibid.

[95] Hermansyah, A. (2017, August 4). Indonesia, Google to use Trusted Flagger program to filter out internet content. Jakarta Post. https://www.thejakartapost.com/news/2017/08/04/indonesia-google-to-use-trusted-flagger-program-to-filter-out-internet-content.html

[96] Tech against terrorism. (n.d.). Trusted Flagger Platform. https://techagainstterrorism.org/tfp

[97] Bernama. (2026).

[98] Clubb, G. (n.d.). Transparency, Evaluation and Standards: Avoiding Pitfalls. Centre for Research and Evidence on Security Threats. https://www.crestresearch.ac.uk/comment/transparency-evaluation-and-standards-avoiding-pitfalls/; and Rogers, E. (2025). The need for greater transparency in the moderation of borderline terrorist and violent extremist content. Internet Policy Review, 14(3). https://doi.org/10.14763/2025.3.2012

[99] MCMC. (2024). Integrated Annual Report 2023: Leading the future, transcending the boundaries of connectivity. https://www.mcmc.gov.my/skmmgovmy/media/General/pdf2/MCMC-IAR-3-2-2024_ENG_Hires_Page-1.pdf

[100] OECD. (2024). Transparency reporting on terrorist and violent extremist content online: Fourth edition (Digital Economy Papers No. 367). https://doi.org/10.1787/901cb8cf-en

[101] Squires, D. (2016). Terrorism and pre-emptive civil processes. The International Journal of Human Rights, 20(5), 684–702. https://doi.org/10.1080/13642987.2016.1162413

[102] Anonymous officers, UNODC, personal communication, February 25, 2026.

[103] Ibid.

[104] Aryaeinejad & Scherer (2024).

[105] Anonymous officers, Wahid Institute, personal communication, February 26, 2026.

[106] Cairney, P. (2019). Understanding Public Policy: Theories and Issues (2nd ed.). Bloomsbury Publishing.

[107] Kumar, A. (2025, October 27). Hybrid moderation models: balancing AI and human oversight. Infosys BPM. https://www.infosysbpm.com/blogs/trust-safety/hybrid-moderation-models-balancing-ai-and-human-oversight.html

[108] Chia, S., Shamsuddin, A., & Azuddin, A. (2025, December). Far-Right Extremism & Tech Accountability in Malaysia (Special Report Issue 3/2025). Initiative to Promote Tolerance and Prevent Violence. https://initiate.my/far-right-extremism-tech-accountability-in-malaysia/

[109] Ingram, M. (2021, March 11). What should we do about the algorithmic amplification of disinformation? Columbia Journalism Review. https://www.cjr.org/the_media_today/what-should-we-do-about-the-algorithmic-amplification-of-disinformation.php

[110] Bernama. (2025, November 25). MCMC remains sole regulater [sic] under Online Safety Act – Azalina. New Straits Times. https://www.nst.com.my/news/nation/2025/11/1322925/mcmc-remains-sole-regulater-under-online-safety-act-azalina?source=widget

[111] Royal Commission of Inquiry (New Zealand). (n.d.) Harmful behaviours, right-wing extremism and radicalisation. https://christchurchattack.royalcommission.nz/the-report/part-2-context/harmful-behaviours-right-wing-extremism-and-radicalisation

cert.isis.org.my