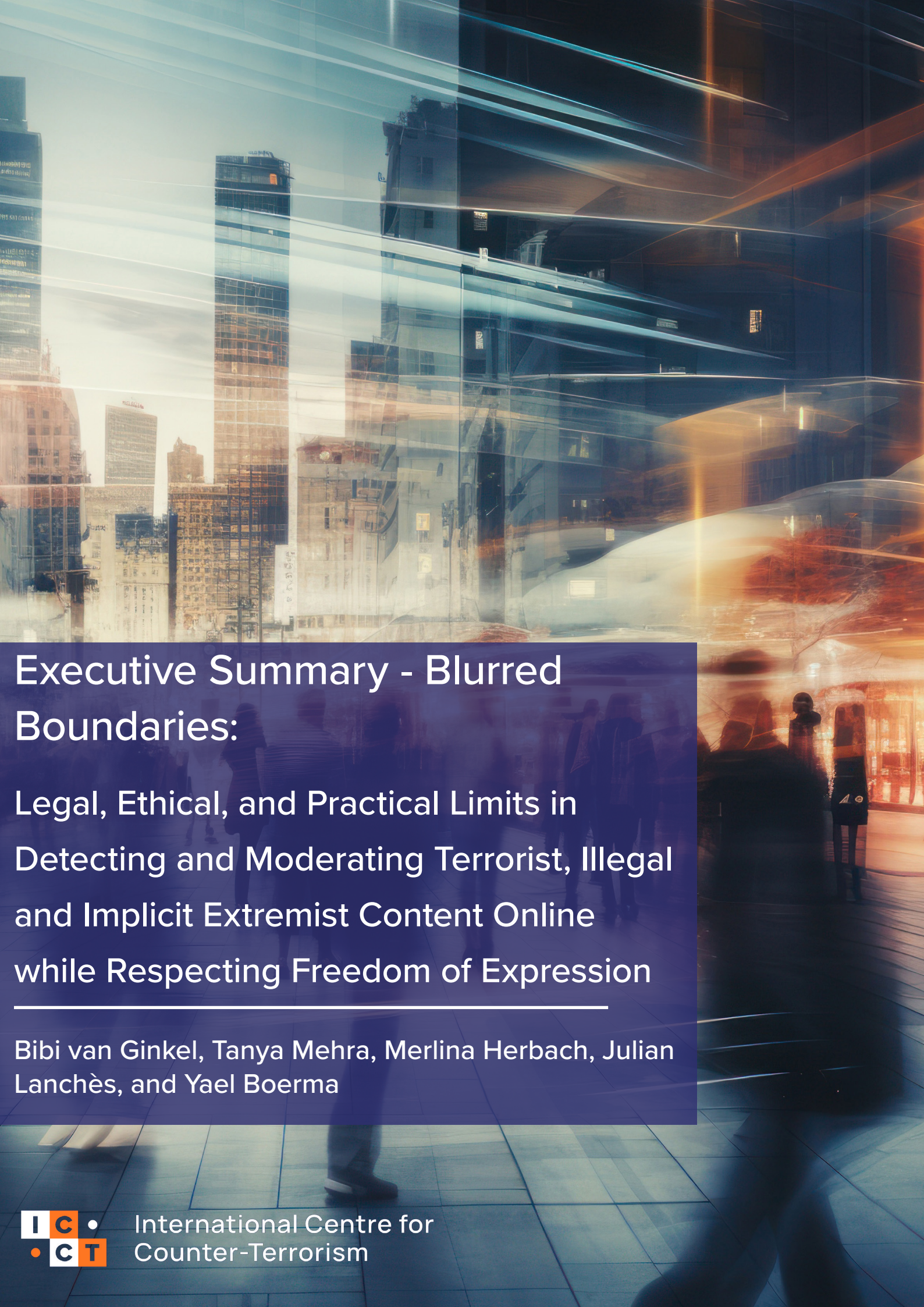

Executive Summary - Blurred Boundaries:

Legal, Ethical, and Practical Limits in Detecting and Moderating Terrorist, Illegal and Implicit Extremist Content Online while Respecting Freedom of Expression

Bibi van Ginkel, Tanya Mehra, Merlina Herbach, Julian Lanchès, and Yael Boerma

International Centre for Counter-Terrorism

Executive Summary - Blurred Boundaries: Legal, Ethical, and Practical Limits in Detecting and Moderating Terrorist, Illegal and Implicit Extremist Content Online while Respecting Freedom of Expression

Bibi van Ginkel, Tanya Mehra, Merlina Herbach, Julian Lanchès, and Yael Boerma

ICCT Report

October 2025

International Centre for Counter-Terrorism

About ICCT

The International Centre for Counter-Terrorism (ICCT) is an independent think and do tank providing multidisciplinary policy advice and practical, solution-oriented implementation support on prevention and the rule of law, two vital pillars of effective counter-terrorism.

ICCT's work focuses on themes at the intersection of countering violent extremism and criminal justice sector responses, as well as human rights-related aspects of counter-terrorism. The major project areas concern countering violent extremism, rule of law, foreign fighters, country and regional analysis, rehabilitation, civil society engagement and victims' voices. Functioning as a nucleus within the international counter-terrorism network, ICCT connects experts, policymakers, civil society actors and practitioners from different fields by providing a platform for productive collaboration, practical analysis, and exchange of experiences and expertise, with the ultimate aim of identifying innovative and comprehensive approaches to preventing and countering terrorism.

Licensing and Distribution

Photocredit: Pravit/Adobe Stock

Introduction and Background

In June 2024, the Dutch Terrorist Threat Assessment (DTN) issued a stark warning: terrorist and extremist actors continue to exploit online platforms to disseminate propaganda, coordinate activities, and incite violence. This content ranges from overtly terrorist material to so-called 'borderline' content - material that does not clearly fall within the legal definitions of terrorist or illegal content but nonetheless exerts corrosive effects on democratic values and social cohesion, therefore also referred to as 'awful but lawful'.

The proliferation of harmful content online is not new, yet its scope, speed, and sophistication have expanded dramatically with technological innovation. The emergence of generative artificial intelligence (AI) and large language models (LLMs) has enabled extremist actors to create and disseminate content at unprecedented volume, speed, and precision, often in multiple languages simultaneously. Jihadist groups such as ISIS and Al-Qaeda, as well as right-wing extremist movements in Europe and North America, already deploy these technologies in their propaganda campaigns.

The stakes are high. Such content can incite hatred, normalise terrorist violence and deepen societal polarisation. Young people, who primarily access information through social media, are particularly vulnerable to online radicalisation. Extremist actors exploit not only mainstream social media platforms, but also gaming environments, streaming sites, and file-sharing networks. The shift from obscure, difficult-to-access corners of the internet a decade ago to openly accessible platforms today underscore the urgency of the problem.

At the same time, the issue raises complex dilemmas at the intersection of security, technology, and fundamental rights. Harmful content often masquerades as humour, irony, or satire, blurring the line between protected freedom of expression and incitement to violence. Overly broad content moderation risks stifling legitimate debate, while insufficient action leaves societies exposed to manipulation and radicalisation.

The role of the technology sector is pivotal in addressing the spread of harmful content online, yet its response has been uneven and increasingly subject to criticism. While platforms bear significant responsibility for detection and moderation, their cooperation with independent research and public institutions remains limited. This reluctance reflects broader concerns that major technology companies prioritise profit-driven strategies over societal responsibility, investing heavily in automated detection systems while simultaneously downsizing teams of human moderators. Such trends risk undermining both the quality and legitimacy of moderation, given that algorithmic tools alone are ill-suited to capture the nuance and context of implicit extremist content.

Against this backdrop, the Research and Data Centre (WODC), at the request of the Netherlands' National Coordinator for Counterterrorism and Security (NCTV), commissioned a study - conducted by ICCT - examining the feasibility of developing a reliable framework for detecting and moderating extremist and terrorist content online, without limiting the freedom of expression. Included in this category of content is the so-called 'borderline' content, which is not always easily detectable due to its implicit character.

The Societal Challenge

Harmful online content poses a profound challenge to democratic, rule-of-law–based societies because it corrodes the very foundations of pluralism, trust, and social cohesion on which they depend. Terrorist propaganda, extremist narratives, and more implicit forms of hateful or divisive speech do not only target individuals or groups; they seek to destabilise democratic institutions by normalising violence, fuelling polarisation, and eroding confidence in the state's ability to protect its citizens. Left unchecked, such content amplifies grievances, deepens societal fractures, and undermines the principles of free and open debate that sustain democratic life. Online platforms are central arenas of modern public life. They host political debates, cultural exchanges, and social interactions. Yet these same spaces are exploited by extremist and terrorist actors, who weaponise communication tools to advance ideological agendas.

This executive summary synthesises the background, research questions, findings, challenges, and recommendations of that study. It provides a critical reflection on the potential and limitations of content detection frameworks and outlines concrete steps for policymakers, online service providers, and other stakeholders. Before doing so, we define the scope of this study and elaborate on the methodology used.

Scope

Considering the ambiguity surrounding existing definitions - most notably the concept of *borderline content* - the research team deemed it necessary to deviate from the regularly used terminology. For the purposes of this study, and to limit the scope, we employ the categories of *terrorist content*, *illegal content*, and *implicit extremist content* that is harmful in the context of violent extremism and terrorism. This choice is especially relevant to the feasibility question of an assessment framework: determining whether content falls under the protection of freedom of expression requires reliance on clear legal definitions. Restrictions on expression are only legitimate when grounded in law and when they meet the criteria of proportionality, necessity, and adequacy. Because neither 'extremist content' nor 'borderline content' is defined in law, they risk being interpreted differently by various stakeholders. We therefore opted for the term *implicit extremist content* to describe material that may not initially appear to fall into the terrorist or illegal categories due to its concealed nature, yet is nonetheless harmful and potentially conducive to radicalisation.

Methodology

This study employed a mixed-methods approach combining desk research, semi-structured interviews, an expert roundtable, and qualitative content analysis. The desk research reviewed academic literature, policy reports, legislation, jurisprudence, and platform Terms of Service up to June 2025. This provided the foundation for identifying indicators relevant to an assessment framework. Semi-structured interviews were conducted with key stakeholders - including government agencies, law enforcement, EU bodies, and NGOs - though platforms declined to participate. Indeed, this study encountered significant barriers to engagement with host service providers, underscoring ongoing concerns about the sector's lack of cooperation. Insights were complemented by a roundtable consultation with experts and practitioners, focusing on definitional challenges, detection and moderation methods, and the feasibility of a structured framework.

To assess the operationalisation of identified indicators, the research team developed and piloted a codebook through a qualitative content analysis of online material collected via open source intelligence (OSINT) research. For this form of qualitative content analysis, we used a contactless

and anonymised scraping method to collect online posts from several accounts. The OSINT research was guided by a data protection impact assessment and strict ethical safeguards. The OSINT contactless research served the purpose of piloting a selection of indicators/markers that could assist in identifying the mentioned content. These indicators were used to develop a pilot assessment framework (hereafter referred to as the pilot codebook) to assist the team in reflecting on the overall feasibility question.

The content scraped in this OSINT phase for this purpose was therefore not used to gain insight into the level, amount, author, or type of content on key platforms. The research question tailored to those aspects will be answered based on the outcomes of the desktop research. Three platforms were selected for the OSINT research - Instagram, TikTok, and Reddit - based on criteria such as relevance to the Dutch context, accessibility for open-source analysis, diversity of user demographics, and variation in moderation practices (e.g., AI-dominant on TikTok and Instagram, mixed human–automated approaches on Reddit).

Content was collected in relation to two nationally significant triggering events. Firstly, the Amsterdam riots in November 2024, when there was a violent confrontation between fans from the football club *Maccabi Tel Aviv*, and a group of people with strong feelings about the Israel-Gaza conflict also expressing anti-Semitic sentiments and chasing and assaulting Maccabi fans,. And secondly the *White Lives Matter* projection on the Erasmus Bridge during New Year's eve for which two observation periods were chosen, namely January 2023 right after the projection, and December 2024 - January 2025 during the court case. Nine accounts were ultimately selected across the three platforms, reflecting both right-wing extremist and Islamist-inspired narratives. All posts published by these accounts within the defined observation windows were manually scraped, anonymised, and stored securely. Posts were then coded using the pilot codebook, which tested indicators across three categories - terrorist, illegal, and implicit extremist content - providing empirical input to assess the feasibility of an assessment framework.

Varieties of Harmful Content

Harmful online content manifests in several overlapping categories. We focus for the purpose of this report on the categories with relevance to terrorism and radicalisation to (violent) extremism:

- **Terrorist content** is defined in accordance with Regulation (EU) 2021/784, which states that terrorist content includes any material that (i) incites or solicits someone to commit or contribute to terrorist offences (ii) Solicits participation in activities of a terrorist group, (iii) glorifies terrorist activities, including by sharing material depicting terrorist attacks, or (iv) provides instructions on making or using explosives, firearms, or other weapons, including chemical, biological, radiological, or nuclear substances. In doing so, terrorist offences are defined pursuant to Article 3 of the Directive (EU) 2017/541.

- **Illegal content** refers to online content that is illegal under national or European law. This includes content that is illegal by itself as well as content that infringes on the consumer protection laws or constitutes a violation of intellectual property rights. For the scope of this study, we will only focus on illegal content in the context of terrorism and violent extremism. This can include hate speech and online content that contributes to polarisation and radicalisation.

- **Implicit extremist content that is harmful:**
  - **'Extremist'** refers to the fact that the content disseminates exclusionary and hateful narratives that may contribute to radicalisation towards terrorism and (violent) extremism.
  - **'Implicit'** refers to the fact that the meaning is concealed. When this is done intentionally, it aims to disguise the illegality, unlawfulness or harmfulness of the content.

- **'Harmful'** refers to the fact that the content could cause serious harm to an individual, a group of people, institutions or to the democratic legal order, and that is not protected under international human rights law.

The category of implicit extremist content is particularly problematic. Its ambiguity shields it from immediate legal sanction while allowing it to sow division and reinforce extremist worldviews. Moreover, extremist actors deliberately calibrate their messaging to remain within the grey zone, ensuring that content evades moderation while still achieving radicalising effects.

Concealment and Adaptation

Extremist actors employ concealment strategies to circumvent detection. Dog whistles, coded emojis, and historical or cultural references intelligible only to in-groups are frequently deployed. Humour and irony, particularly through memes, serve both as rhetorical shields and as recruitment tools, normalising extremist ideas while deflecting external criticism.

These actors are agile and adaptive. As platforms strengthen moderation of overtly terrorist content, extremists shift toward more implicit forms, carefully crafting their discourse to appear 'awful but lawful'. Emerging technologies amplify this trend: generative AI facilitates the production of sophisticated text, images, and videos, while deepfake technologies and interactive gaming environments provide new accelerating instruments for dissemination.

Societal Impacts

The consequences of unchecked harmful content are significant:

- **Radicalisation pathways**: exposure to extremist narratives online is a well-documented factor towards violent radicalisation, particularly among youth.
- **Normalisation of violence**: repeated exposure to extremist rhetoric reduces social resistance to violence, embedding extremist worldviews in mainstream discourse.
- **Polarisation**: harmful content deepens societal divisions, erodes trust in institutions, and fuels hostility between communities.
- **Democratic resilience**: the manipulation of online discourse undermines open democratic debate, narrowing the space for pluralism and constructive disagreement.

The societal challenge is therefore twofold: preventing the exploitation of online platforms for extremist purposes while safeguarding fundamental freedoms, including the right to freedom of expression.

The Regulatory and Institutional Landscape

European Frameworks

In response to these threats, the European Union has developed a layered regulatory framework. Two instruments stand out:

- **Regulation on Dissemination of Terrorist Content Online (TCO, Regulation EU 2021/784)**: obliges hosting service providers (HSPs) to remove terrorist content within one hour of notification by competent authorities. It introduces transparency requirements, user notification, and differentiated obligations depending on exposure levels. Critics highlight the disproportionate burden on smaller platforms lacking resources to comply.
- **Digital Services Act (DSA, Regulation EU 2022/2065)**: significantly broadens the regulatory scope to cover all illegal content and imposes obligations on very large online platforms (VLOPs) and search engines (VLOSEs) to conduct systemic risk assessments. It enhances transparency by requiring annual reports, creating a DSA Transparency Database, and granting researchers access to platform data.

Together, these frameworks mark a shift from voluntary cooperation with platforms toward a regulatory model that seeks harmonisation, accountability, and minimum human rights safeguards. Yet implementation challenges remain, particularly concerning proportionality, capacity constraints of smaller providers, and the protection of fundamental rights.

Platform Governance

Alongside legal frameworks, platforms enforce their own Terms of Service (ToS) and community guidelines. These rules often extend beyond legal obligations, encompassing broader categories of harmful content. While this proactive stance may limit regulatory fines, it raises concerns about private actors effectively setting the boundaries of online freedom of expression without clear democratic oversight.

A lack of transparency compounds the problem. Users frequently struggle to understand why content is removed, while researchers and regulators face obstacles in accessing moderation data. The opacity of ToS enforcement decisions undermines trust and accountability.

Fundamental Rights Dimension

The right to freedom of expression is at the core of this debate. Content moderation inevitably involves normative judgments about what is permissible. When platforms err on the side of caution, they risk removing legitimate critique, satire, or dissent, with disproportionate effects on marginalised or minority groups. Conversely, insufficient moderation allows harmful narratives to flourish unchecked. Courts remain the ultimate arbiters in disputes, but litigation is slow and rare, offering little timely guidance.

The regulatory challenge, therefore, is not only about ensuring compliance but also about embedding fundamental rights safeguards into moderation practices.

Research Questions

The commissioned study set out to examine whether a reliable assessment framework could be developed to help platforms identify and moderate terrorist, illegal, and implicit extremist content. The research was guided by seven core questions:

1. What features determine whether online content constitutes terrorist, illegal, or implicit extremist material?
2. How can these features be detected and identified on online platforms?
3. How reliable are current detection methods for these categories?
4. How does detection reliability relate to risks of wrongful moderation and infringements on freedom of expression?
5. Is it possible, with current knowledge and technologies, to develop a valid and reliable interpretive framework for detecting harmful content without unjustly infringing rights?
6. Under what conditions could such a framework contribute to reducing harmful content and online radicalisation?
7. If not feasible, what barriers prevent the development of such a framework, and how might they be overcome?

These questions provided the analytical lens for assessing feasibility, reliability, and ethical desirability.

Key Findings

Based on the desktop research, the team developed a pilot codebook with key indicators to qualify terrorist, illegal, and implicit extremist content online (research questions 1 and 2). Since the study's objective was not to design a full framework but to test its feasibility, the scope of the pilot codebook was deliberately limited. To keep the exercise manageable, given the labour-intensive coding process, only a selection of indicators was included. The focus was on right-wing extremist, jihadist, and implicit extremist content. For terrorist content, the team prioritised types most open to dispute, while for illegal content, only forms with potential overlap with implicit extremist content were considered.

For the qualification of terrorist content, the pilot focused on three crime types: incitement to commit or participate in a terrorist offence, glorification of terrorist acts, and recruitment to a terrorist organisation. The indicators were drawn from the EU Terrorist Content Online Regulation (EU 2021/784), which is binding on both platforms and competent authorities across the EU. Each post was screened against these criteria to assess whether it could be qualified as terrorist content.

For the qualification of illegal content, only a limited set of non-terrorist categories was included to test the framework's feasibility: hate speech, incitement to violence, and denial, downplaying or justification of international crimes. These were selected based on their potential overlap with extremist narratives and assessed using indicators derived from the Dutch Criminal Code.

In both cases, an additional safeguard ensured that content falling within the scope of freedom of expression was not misclassified as terrorist or illegal.

Indicators of Implicit Extremist Content

While for the previous categories, the research team could refer to legal frameworks to identify the key indicators, for the identification of implicit extremist content, this was not the case. The study, therefore, identified several indicators relevant for classifying implicit extremist content based on the analysis of literature and policy reports. The categories of indicators proposed to identify implicit extremist content are:

- **Concealment of Meaning (CM)**: deliberate obfuscation through irony, humour, or coded language.
- **Harmful Alliances/Affiliations (HA)**: references to extremist groups, ideologies, or symbols.
- **Problematic References (PR)**: invocation of historical or current events with extremist framing.
- **Implicit Action Triggers (AT)**: subtle cues encouraging audiences to take action.
- **Presumed Intent to Cause Harm (IH)**: inferred harmful intent underlying the content.

Each of these categories consisted of several separate indicators. A combination of indicators would need to be present in the content to qualify as implicit extremist content. Here as well, additional safeguard questions were built in to ensure that content falling under the protection of freedom of expression would be exempted. These questions tested whether the content is not satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation.

The indicators of terrorist, illegal and implicit extremist content were incorporated into a pilot codebook to test their reliability (research questions 3 and 4). Indicators were assessed on three criteria: usability, usefulness, and interpretability. *Usability* reflects how easy and time-efficient it is to code an indicator, with green for easy, orange for moderately demanding, and red for difficult indicators. *Usefulness* indicates whether an indicator meaningfully contributes to the overall assessment; red suggests it is redundant, orange signals limited contribution or overlap. *Interpretability* evaluates the subjectivity of coding, with green for clear indicators, orange for moderately subjective ones needing refinement, and red for highly subjective indicators identified through intercoder mismatches.

The study subsequently reflects on the outcomes of the scoring exercise, which yielded the following findings:

Importance of Context

Indicators rarely function in isolation. Interpretation depends heavily on context, including the identity of the speaker, cultural references, and audience reception. This context dependence complicates scaling detection mechanisms and undermines reliability when relying solely on automated systems.

Hybrid Detection Models

The study underscores that automated tools alone cannot capture the nuance of implicit extremist content. A hybrid model - combining AI-assisted pre-screening with human expertise - is essential. Human coders bring contextual sensitivity, but require structured guidance, training, and safeguards against bias.

Operational choices of the Tech sector

Platforms increasingly rely on automated tools as the backbone of their moderation systems, presenting this as a solution to issues of scale and efficiency. Yet the study shows that automation, while useful for detecting overtly terrorist or illegal content, performs poorly when applied to implicit extremist material, which often links to cultural nuances, irony, or coded references. The reduction of human moderators across major platforms exacerbates this risk, creating a growing gap between the complexity of harmful content and the sector's chosen methods for addressing it.

Transparency and Accountability

Platforms' lack of transparency about moderation practices severely limits public trust and academic scrutiny. There is a range of moderation options available relating to the content or the account, including reducing visibility. Each of these options has a different impact on the freedom of expression. It remains unclear, however, how often which moderation decision is taken and what impact the decision has on the freedom of speech (proportionality requirement). Without clear reporting and accessible appeals, users remain in the dark about the rules governing the moderation of their content.

Feasibility of an assessment framework?

Based on the findings, several fundamental challenges were identified. Together, these challenges suggest that while incremental improvements are possible, a universal and fully reliable assessment framework remains infeasible at present.

Challenges Identified

In relation to the feasibility of a reliable assessment framework, the following entrenched obstacles were highlighted:

- **Definitional Ambiguity and Blurry Boundaries**
  - Lack of universally accepted legal definitions of terms like *terrorism*, *hate speech*, *violent extremism* or *incitement*.
  - Lack of internationally accepted definitions for key concepts such as *group*, *legitimate*, or *self-defence*.
  - Unclear thresholds for when content qualifies as harmful versus acceptable.
  - Difficulties distinguishing extremist rhetoric from satire, political critique, or legitimate debate.
  - Vague categories undermine coder consistency and inter-coder agreement.
  - Implicitness can refer to concealing the harmfulness, lawfulness or legality of the content. Understanding when this is done intentionally is difficult.

- **Evolving Concealment Tactics**
  - Extremist actors adapt language and strategies quickly, leaving static codebooks outdated.
  - Dog whistles, irony, memes, and coded terms evade detection.

- **High Subjectivity and Risk of Bias when relying on human assessment**
  - Heavy reliance on personal interpretation and context leads to inconsistent results.
  - Risk of misclassifying ordinary expressions of grievance or opposition as extremist.
  - Coders' cultural, ideological, or political backgrounds can skew judgments.

- Indicators are vulnerable to misuse by biased coders or flaggers.

- **Accuracy v. bias:**
  - Automated moderation might be implementing the instructed algorithm accurately, yet there is a risk of a built-in bias that does not manifest itself quickly.
  - While automated moderation may be cheaper and faster, it might not be able to detect implicit extremist content accurately, yet while human assessment might be better in interpreting implicit extremist content, it is more costly and runs the risk of bias.
  - The sheer volume of online content exceeds the capacity of both human coders and current AI systems.

- **Context Dependence and Labour Intensity**
  - Many indicators require deep knowledge of user history, ideology, or platform dynamics. This reduces usability and makes identification difficult.

A further structural challenge lies in the underlying economic business models that shape the practices of hosting service providers. Driven primarily by profit motives, large technology companies have little incentive to invest in resource-intensive, human-led moderation practices that would improve reliability and safeguard fundamental rights. This research team, like many others, encountered significant reluctance from platforms to cooperate or provide transparency on their moderation approaches, highlighting an accountability gap between private governance and public interest. The downsizing of human moderation teams, combined with the opacity of algorithmic decision-making, not only undermines the reliability of detection but also limits opportunities for democratic oversight.

Recommendations

A universal, valid, and reliable assessment framework for detecting implicit extremist content does not appear feasible (research questions 5-7). However, more reliable and adaptable frameworks used by hosting service providers (HSPs) might be achievable. By refining definitions, operationalising complex indicators, and embedding iterative learning, hybrid systems, and collaborative oversight, the indicators used can be transformed into more consistent, useful and reliable tools. Such an approach balances the need to identify implicit extremist content with safeguards that protect freedom of expression. Ultimately, and as long as there are no additional legal frameworks applicable, the effectiveness of any detection mechanism will depend on the willingness and capacity of HSPs to apply it responsibly. Governments, together with the EU, should intensify their dialogue with HSPs to stimulate this process.

Whether such a universal assessment framework is even desirable remains an open question. Ethical considerations must play a central role in shaping the way forward. More importantly, contemporary communication is increasingly complex and multi-layered: online and offline spheres are deeply intertwined, reflecting both social diversity and growing polarisation. This rapid transformation has outpaced public debate on the norms and etiquette of online communication. Especially when legal frameworks are ambiguous or inconsistent in setting boundaries, the development of any assessment framework for online content should begin with broad societal dialogue about what constitutes acceptable expression and what crosses into unacceptably harmful territory.

Despite these reservations, below we highlight recommendations for the Tech sector and for the policymakers. The recommendations mostly follow directly from the findings in this study, some are, however, derived from the general expertise and practical experiences the research team members have in implementing prevention programmes and capacity-building.

Key conditions HSPs can implement for a reliable and accurate assessment framework:

1. **Clear Definitions and Thresholds**
   a. Abide by the obligations to implement precise definitions for vague terms;
   b. Establish thresholds for incitement, hate, or hostility that do not restrict the freedom of expression;
   c. Be clear on the combination of indicators that need to be met to qualify content as implicit extremist content.

2. **Guidance and Examples**
   a. Offer illustrative examples across ideological spectrums, including examples of when content is protected by the freedom of expression, such as satire, critique, and harmful but lawful content;
   b. Use decision-making frameworks or coding flowcharts to standardise application;
   c. Develop typologies of in-/out-groups to guide application.
   d. Provide coder training, prompts, and bias-mitigation strategies;
   e. Adopt "four-eyes" review protocols and consensus-building practices for borderline cases.

3. **Complex Indicators Operationalisation**
   a. Break broad indicators into sub-categories or spectrum-based assessments;
   b. Use triangulation with contextual clues (e.g., history, platform dynamics) to improve reliability.

4. **Iterative and Adaptive Frameworks**
   a. Tailor frameworks to particular extremist narratives (e.g., jihadist, right-wing extremism, separatist) and cultural contexts and avoid a one-size-fits-all approach;
   b. Treat assessment frameworks as "living documents" that evolve with new extremist tactics;
   c. Regularly update with fresh examples from monitoring and research.

5. **Hybrid AI–Human Systems**
   a. Combine AI pre-screening for scale with expert human review for contextual judgment;
   b. Use LLMs and databanks to handle labour-intensive tasks while keeping humans in the loop.

6. **Multi-Stakeholder Collaboration**
   a. Regularly engage with researchers, practitioners, platforms, civil society, and affected communities in the development and revision of assessment frameworks;
   b. Build a broad consensus to reduce risks of bias, overreach, and misclassification.

7. **Transparency and Appeal**
   a. Improve transparency and reporting on moderation decisions;
   b. Need for respect of the proportionality principle regarding moderation decisions to better ensure freedom of expression: To ensure that the freedom of expression is respected, a broader range of proportional moderation decisions needs to be developed, but also implemented in practice;
   c. Provide clear information on how and where a moderation decision can be appealed.

Specific Recommendations for policymakers:

This study was conducted at the request of NCTV. Although the findings and recommendations of this study are relevant for a broader audience, the NCTV is one of the key coordinating actors that plays a key role in the Netherlands in shaping and implementing policies, engaging with other European partners in furthering European policies, and in the dialogue with HSPs. Based on the findings in this study, we formulated tailored recommendations for policymakers, in particular those in a coordinating role, such as the NCTV and the ATKM.

Policymakers are recommended to:

1. **Refrain from using the term 'borderline' content**, as that will only further adhere to the confusion about the scope and meaning of the term.

2. **Set up a multistakeholder group**, consisting of researchers, practitioners, platforms, civil society, and affected communities, to (regularly) and transparently reflect on, and publicly report on:
   a. A set of indicators to detect terrorist, illegal and implicit extremist content, in line with the idea of maintaining a living document;
   b. The indicators that, according to this study, are deemed problematic and to improve their formulation;
   c. A threshold to be used in the combination of indicators to detect implicit extremist content;
   d. Key incidents or historical facts specific to the Netherlands, as well as key expressions, language used**,** prompts or codes specific to the Dutch language by extremist groups active online in the Netherlands, that can assist in the contextual interpretation of implicit extremist content.

3. **Promote public debate**
   a. On what is harmful and unlawful, and what is harmful but lawful content;
   b. On how much autonomy HSPs should have in facilitating public space for public speech.

4. **Support media literacy training in schools, prevention programmes for youth organisations and the use of strategic de-escalation communicative engagement techniques** to confront or debate the use of particular harmful content. Offer these trainings also to minority and marginalised groups to build resilience.

5. **Develop a clear strategic communication policy** on how to respond to harmful but lawful content, and explain why something is considered harmful. Meanwhile, also **speak up clearly against harmful, unlawful content,** especially when that targets minority or marginalised groups. Also offer guidance to local policymakers on strategic communication.

The coordinating government actors (such as NCTV and ATKM) are recommended to entertain a transparent and open dialogue with the HSP

6. **To (continue to) engage, for the purpose of transparency, with big and small online service providers:**
   a. To conduct an open discussion on the indicators they use in their assessment frameworks and whether they use different assessment frameworks for different ideologies;
   b. To enhance information exchange and transparency on ways to proportionally moderate content in line with the freedom of expression.

7. Without releasing HSPs of their primary responsibility, yet considering Dutch is a small language, **to share the list of expressions, language used, prompts or codes specific to**

the Dutch language used by extremist groups active online in the Netherlands, that was discussed in the multistakeholder group, to assist in the contextual interpretation of implicit extremist content.

8. **To regularly provide contextual background briefs** to educate HSP on typically Dutch (topical or historical) events, which can assist them with the contextual interpretation of online content.

The coordinating government actors are recommended to engage in dialogue with other European Member States and the European Commission:

9. **To cooperate with the sector to develop a sector wide code of conduct,** which offers a certification ('keurmerk') that offers consumers a better understanding of how HSPs conduct their detection and moderation; setting standards for the percentage of human assessment, clarity on the terms used in the ToS, filters implemented to protect vulnerable groups, transparency on moderation decisions, and appeals procedures.

10. **To engage in further strengthening the regulatory frameworks** demanding more transparency and accountability of HSPs, demanding an ex-ante evaluation on how they respect the freedom of expression by applying their ToS, regular ex-post evaluations of how freedom of expression was respected in moderation decisions, stricter rules on moderation methods (AI versus manual), and by providing clear definitions and guidance.

Conclusion

The study concludes that developing a universal, reliable assessment framework for detecting terrorist, illegal, and implicit extremist content online is not feasible under current legal, technical, and ethical conditions. Definitional ambiguity, contextual complexity, and the evolving tactics of extremist actors undermine the reliability of such frameworks and heighten risks to freedom of expression.

Yet the infeasibility of a universal assessment framework does not mean progress is impossible. Incremental improvements are achievable through clearer definitions, adaptive indicators, hybrid AI–human models, and multi-stakeholder collaboration. Rather than seeking a one-size-fits-all solution, platforms and policymakers should pursue flexible, evolving approaches tailored to specific contexts and ideologies.

For the NCTV and the wider policy community, the challenge extends beyond detection to broader societal questions. What constitutes harmful but lawful content? How much autonomy should private companies have in governing online public spaces? How can societies strike a balance between security and freedom?

Ultimately, the path forward lies in continuous dialogue, transparency, and adaptability. As extremist actors innovate, so too must policymakers, platforms, and communities. Building resilience requires not only technological solutions but also democratic debate about the norms of online communication and the values societies wish to uphold.

International Centre for Counter-Terrorism

International Centre for Counter-Terrorism (ICCT)
T: +31 (0)70 763 0050
E: m.e.centre@icct.nl
https://icct.nl/Monitoring-and-Evaluation