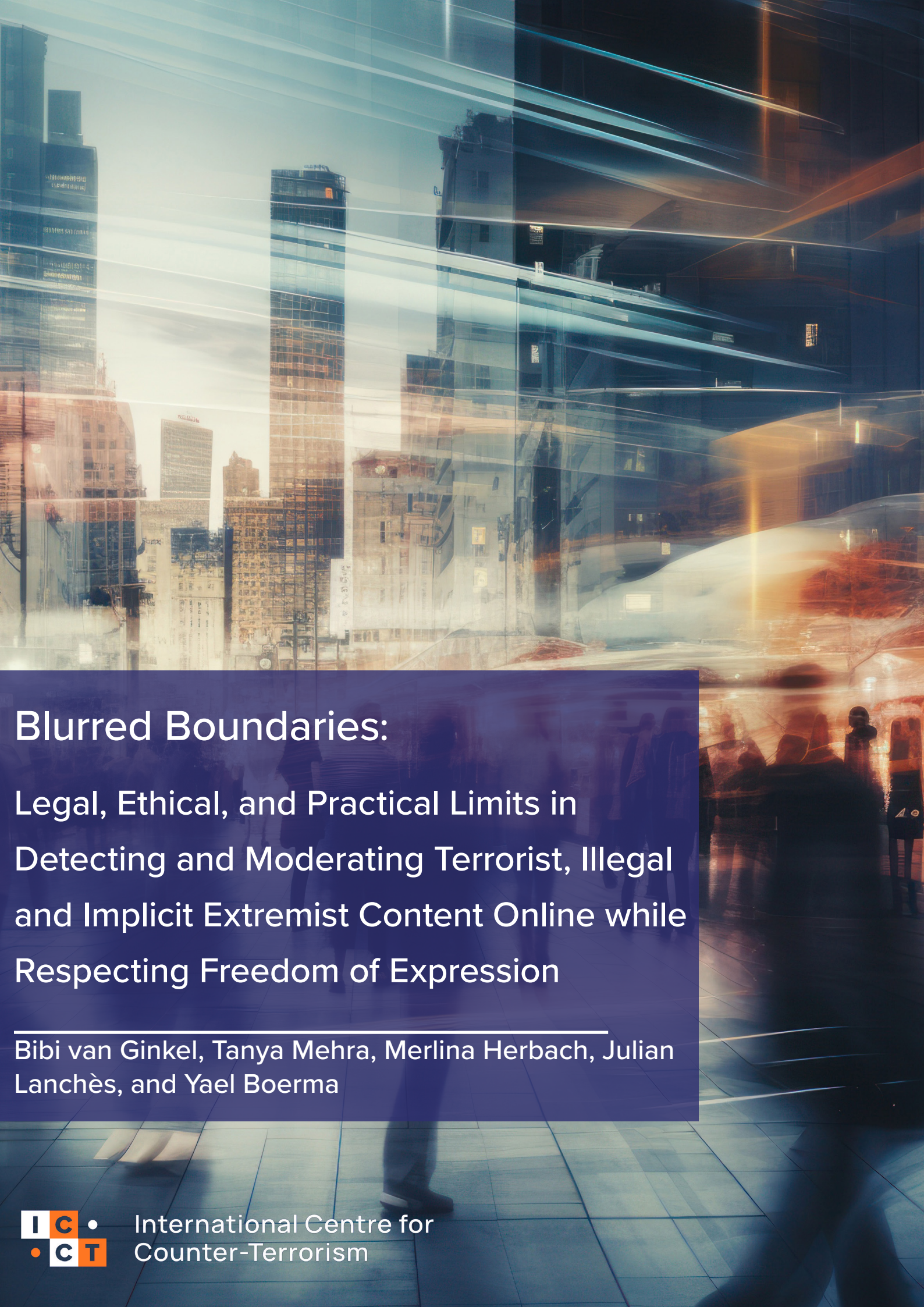


Blurred Boundaries:

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

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

November 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

Foreword

This study examines a pressing and highly topical challenge: how to assess online content that may undermine democracy, threaten national security and public safety, or infringe upon the rights of others—while safeguarding freedom of expression. The central question it explores, the specific challenges identified, and the recommendations it puts forward should not be viewed in a vacuum. Rather, they are situated within a broader and increasingly complex societal and political context. A range of systemic developments shapes the environment in which this work takes place: the rise of online radicalisation, particularly among children and young adults; the expanding influence of large technology platforms and the tensions this creates with rule-of-law-based democratic societies leading to a global trend toward both techno-libertarianism and techno-authoritarianism; and the evolving role of governments as they seek to reconcile the imperatives of security, safety, and national interest with those of privacy, human rights, and minority protection. These challenges are compounded by the unprecedented speed and scale of online information dissemination, growing concerns about disinformation and foreign influence, and the urgent need to strengthen societal resilience and media literacy. While this study does not address each of these systemic issues in depth, they form the essential backdrop against which its findings and proposals should be understood.

Acknowledgements

We would like to thank WODC for granting us the opportunity to conduct this study. Special thanks to the members of the Advisory Committee, who offered welcome feedback and suggestions, sharpening our methodological approach.[1] We are furthermore grateful to the participants of the expert round table and everyone that granted us an interview, both of which provided crucial insights that assisted us in reflecting on the key research questions. This research has been a group effort of an exceptionally dedicated team. Crucially important to the work have been the contributions of Ms. Méryl Demuynck, Msc, former Senior Researcher of ICCT, to the development of the pilot codebook, Ms. Juliette Masclet, LLM, ICCT research assistant for her assistance in the literature review, and Ms Noemie Abtan on the editing and formatting of the report.

On behalf of the ICCT Research Team,

Dr. Bibi van Ginkel, LLM, Project coordinator

1 The Advisory Committee consisted of the following members: Prof. dr. Richard Rogers (University of Amsterdam, chair), dr. H.C.J. van der Veen (WODC), drs. R. van Wonderen (Verweij-Jonker Instituut), mw. dr A.J. Leidinger (University of Amsterdam), prof. mr. dr. L.A. van Noorloos (University of Leiden), mr. P.J. Leerssen, LLM (University of Amsterdam), a representative of NCTV.

Contents

List of Abbreviations

AfD	Alternative for Germany (Alternative für Deutschland)
AI	Artificial intelligence
AI	Artificial Intelligence Act
AIVD	General Intelligence and Security Service (Algemene Inlichtingen- en Veiligheidsdienst)
API	Application programming interface
AT	Implicit action trigger(s)
ATKM	Authority for the Prevention of Online Terrorist Content and Child Sexual Abuse MaterialAuthority for Online Terrorist and Child Pornographic Material (Autoriteit online Terroristisch en Kinderpornografisch Materiaal)
AVMSD	Audiovisual Media Services Directive
BKA	Federal Criminal Police Office (Bundeskriminalamt)
BMI	Federal Ministry of the Interior (Bundesministerium des Innern)
CDMSI	Steering Committee for Media and Information Society
CG	Community Guidelines
CM	Intent to conceal meaning
CSAM	Child sexual abuse material
EU DG HOME	European Union Directorate-General for Migration and Home Affairs
DIC	Denial, downplaying or justification of international crimes
DPIA	Data protection impact assessment
DSA	Digital Services Act
DTN	Terrorist Threat Assessment Netherlands (Dreigingsbeeld Terrorisme Nederland)
EC	European Commission
ECTC	European Counter Terrorism Centre
ECtHR	European Court of Human Rights
EDPS	European Data Protection Supervisor
EIC	Explicit illegal content
EU	European Union
EU IRU	European Union Internet Referral Unit
FRISCO	Fighting Terrorist Content Online
FTO	Foreign terrorist organisation
GDPR	General Data Protection Regulation
GIFCT	GLobal Internet Forum to Counter Terrorism
GT	Glorification of a terrorist act
HA	Harmful alliances/affiliations
HS	Hate speech
HSP	Hosting service provider
ICCT	International Centre for COunter-Terrorism
IH	Presumed intent to cause harm
IHC	Implicit extremist content that is potentially harmful

IRU	Internet Referral Unit
ISIS	Islamic State
ISKP	Islamic State - Khorosan Province
IT	Incitement to commit/participate in a terrorist offence
IV	Incitement to violence
JHA	Justice and Home Affairs
LGBTQIA+	Lesbian, gay, bisexual, transgender, queer, intersex, asexual and more
LLM	Large language model
MS	Member State
NCTV	National Coordinator for Counterterrorism and Security (Nationaal Coördinator Terrorismebestrijding en Veiligheid)
NetzDG	Network Enforcement Act (Netzwerkdurchsetzungsgesetz)
NGO	Non-governmental organisation
NRM	Nordic Resistance Movement
NSFW	Not suitable for work
OSCE	Organisation for Cooperation and Security in Europe
OSINT	Open source intelligence
PA	Patriotic Alternative
PERCI	Plateforme Européenne de Retraits de Contenus Illégaux sur Internet
PR	Problematic reference(s) to historical/current context
RAN	Radicalisation Awareness Network
RT	Recruitment for a terrorist organisation
RWX	Right-wing extremism
SDGT	Specially designated global terrorist
SDNTK	Specially desugnated narcotics trafficking kingpin
SoR	Statement of Reasons
TATE	Tech Against Terrorism Europe
TC	Terrorist Content
TCO	Terrorist Content Online Regulation
TSPA	Trust and Safety Professionals Association
TVE	Terrorism and violent extremism
ToS	Terms of Service
ToU	Terms of Use
UEFA	Union of European Football Association
UK	United Kingdom
UN	United Nations
UNGP	United Nations Guiding Principles on Business and Human Rights
URL	Uniform Resource Locator
US	United States
VE	Violent extremism
VLOSE	Very large online search engine
VLOP	Very large online platform
WODC	Research and Data Centre (Wetenschappelijk Onderzoek en Datacentrum)

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.

Inleiding en Achtergrond

In juni 2024 gaf het Dreigingsbeeld Terrorisme Nederland (DTN) een scherpe waarschuwing af: terroristische en extremistische actoren blijven onlineplatforms misbruiken om propaganda te verspreiden, activiteiten te coördineren en tot geweld aan te zetten. Deze content varieert van openlijk terroristisch materiaal tot zogenoemde *'borderline'*-content - materiaal dat niet duidelijk binnen de wettelijke definities van terroristische of illegale content valt, maar desalniettemin een erosief effect heeft op democratische waarden en sociale cohesie. Daarom wordt dit ook wel aangeduid als *'awful but lawful'* ('afschuwelijk maar wettelijk toegestaan').

De online verspreiding van schadelijke content is niet nieuw, maar de reikwijdte, snelheid en verfijning ervan zijn door technologische innovaties drastisch toegenomen. De opkomst van generatieve kunstmatige intelligentie (AI) en grote taalmodellen (LLM's) stelt extremistische actoren in staat om content te creëren en te verspreiden op ongekende schaal, snelheid en precisie, vaak in meerdere talen tegelijk. Jihadistische groeperingen zoals ISIS en Al-Qaida, evenals rechts-extremistische groeperingen in Europa en Noord-Amerika, maken reeds gebruik van deze technologieën in hun propagandacampagnes.

De belangen zijn groot. Dergelijke content kan haat aanwakkeren, terroristisch geweld normaliseren en maatschappelijke polarisatie verdiepen. Jongeren, die hun informatie voornamelijk via sociale media verkrijgen, zijn bijzonder kwetsbaar voor online radicalisering. Extremistische actoren misbruiken niet alleen reguliere sociale mediaplatforms, maar ook gamingplatforms, streamingwebsites en *file sharing* omgevingen. De verschuiving van deze content van obscure, moeilijk toegankelijke uithoeken van het internet tien jaar geleden naar openbaar toegankelijke platforms vandaag de dag benadrukt de urgentie van dit probleem.

Tegelijkertijd roept dit vraagstuk complexe dilemma's op die op het snijvlak van veiligheid, technologie en fundamentele rechten liggen. Schadelijke content wordt vaak verhuld met humor, ironie of satire, waardoor de grens tussen vrijheid van meningsuiting en aanzetten tot geweld vervaagt. Te brede contentmoderatie kan legitiem debat onderdrukken, terwijl ontoereikende moderatie samenlevingen blootstelt aan manipulatie en radicalisering.

De rol van de technologiesector is cruciaal bij het aanpakken van de verspreiding van schadelijke content, maar hun aanpak is niet eensluidend en staat steeds vaker ter discussie. Hoewel platforms een grote verantwoordelijkheid dragen voor detectie en moderatie, blijft hun samenwerking met onafhankelijk onderzoek en publieke instellingen beperkt. Deze terughoudendheid weerspiegelt bredere zorgen dat grote technologiebedrijven prioriteit geven aan winstgevendheid van bedrijfsvoering ten koste van maatschappelijke verantwoordelijkheid. Ze investeren fors in geautomatiseerde detectiesystemen, terwijl teams van menselijke moderatoren tegelijkertijd worden ingekrompen. Dergelijke trends dreigen zowel de kwaliteit als de legitimiteit van moderatie te ondermijnen, aangezien algoritmische hulpmiddelen op zichzelf vaak ongeschikt zijn om de nuance en context van impliciete extremistische content te herkennen.

Tegen deze achtergrond gaf het Wetenschappelijk Onderzoek- en Datacentrum (WODC) op verzoek van de Nationaal Coördinator Terrorismebestrijding en Veiligheid (NCTV) opdracht tot een studie - uitgevoerd door ICCT - naar de haalbaarheid van het opstellen van een duidingskader dat een betrouwbare detectie en moderatie van extremistische en terroristische, online content mogelijk maakt, zonder dat dit ten koste gaat van de vrijheid van meningsuiting. Onderdeel van de categorie problematische content, is de zogenoemde 'borderline content', die vaak een impliciet karakter heeft, en daardoor niet altijd gemakkelijk te detecteren is.

De Maatschappelijke Uitdaging

De grote hoeveelheid schadelijke content die online beschikbaar is stelt democratische rechtsstaten voor grote uitdagingen. Dergelijke content erodeert immers pluralisme, vertrouwen en sociale cohesie, waarden die het fundament vormen van een democratische rechtsstaat. Terroristische propaganda, extremistische narratieven en meer impliciete vormen van haatdragende of polariserende taal richten zich niet alleen op individuen of groepen; ze beogen democratische instituties te destabiliseren door geweld te normaliseren, polarisatie te voeden en het vertrouwen in het vermogen van de staat om burgers te beschermen te ondermijnen. Zonder toezicht kan dergelijke content gevoelens van onvrede versterken, maatschappelijke breuklijnen verdiepen en de principes van vrij en open debat die essentieel zijn in een democratische samenleving ondermijnen. Onlineplatforms zijn de hedendaagse *agora,* het dorpsplein van het moderne publieke leven. Ze bieden ruimte voor politiek debat, culturele uitwisseling en sociale interacties. Echter, deze ruimtes worden ook gebruikt door extremistische en terroristische actoren die communicatiemiddelen misbruiken om ideologische agenda's te bevorderen.

Deze managementsamenvatting bundelt de achtergrond, onderzoeksvragen, bevindingen, uitdagingen en aanbevelingen van dat onderzoek. Ze biedt een kritische reflectie op de mogelijkheden en beperkingen van contentdetectiekaders en geeft concrete aanbevelingen aan beleidsmakers, online dienstverleners en andere belanghebbenden. Vooraf definiëren we de reikwijdte van het onderzoek en lichten we de gebruikte methodologie toe.

Reikwijdte

Gezien de onduidelijkheid rond bestaande definities - met name het concept *borderline content* - achtte het onderzoeksteam het noodzakelijk om af te wijken van de veel gebruikte terminologie. Voor de doeleinden van dit onderzoek en om de reikwijdte te beperken, hanteren we de categorieën *terroristische content*, *illegale content* en *impliciete extremistische content* die schadelijk is in de context van gewelddadig extremisme en terrorisme. Deze keuze is vooral relevant voor de haalbaarheidsvraag aangaande een duidingskader. Immers, vaststellen of content valt onder de vrijheid van meningsuiting vereist het gebruik van duidelijke juridische definities. Beperkingen op uitingen zijn alleen legitiem wanneer ze in de wet zijn verankerd en voldoen aan de criteria van proportionaliteit, noodzakelijkheid en adequaatheid. Omdat noch 'extremistische content' noch 'borderline content' juridisch is gedefinieerd, bestaat het risico dat verschillende belanghebbenden dit uiteenlopend interpreteren. Daarom kozen we voor de term *impliciete extremistische content* om materiaal te beschrijven dat op het eerste gezicht niet als terroristisch of illegaal herkenbaar is vanwege de verhullende aard, maar toch schadelijk is en potentieel bevorderlijk is voor radicalisering.

Methodologie

Dit onderzoek maakte gebruik van een *mixed-methods* aanpak, waarbij literatuuronderzoek, semigestructureerde interviews, een expertbijeenkomst en kwalitatieve inhoudsanalyse werden gecombineerd. Het literatuuronderzoek omvatte academische literatuur, beleidsrapporten, wetgeving, jurisprudentie en servicevoorwaarden van platforms gepubliceerd voor juni 2025. Dit vormde de basis voor het identificeren van indicatoren die relevant zijn voor een duidingskader. Semigestructureerde interviews werden afgenomen met relevante belanghebbenden - waaronder overheidsinstanties, wetshandhavingsinstanties, EU-organen en ngo's - al weigerden platforms mee te werken. Het gebrek aan bereidheid mee te werken aan dit onderzoek, illustreert de aanhoudende zorgen over het algehele gebrek aan medewerking en transparantie in de sector. De inzichten werden aangevuld met een rondetafelgesprek met experts en praktijkdeskundigen, gericht op definitie-uitdagingen, detectie- en moderatiemethoden en de haalbaarheid van een duidingskader.

Om de operationalisering van geïdentificeerde indicatoren te toetsen, ontwikkelde het onderzoeksteam een codeboek dat werd getest door middel van een kwalitatieve inhoudsanalyse van online data, verzameld via *open source intelligence* (OSINT) onderzoek. Voor deze vorm van een kwalitatieve inhoudsanalyse gebruikten wij een contactloze en geanonimiseerde scrape methode om online content gepost op verschillende accounts te verzamelen. Voorafgaand aan het OSINT onderzoek, is een *data protection impact assessment* uitgevoerd en bovendien werden strikte ethische waarborgen in acht genomen. Het doel van het OSINT onderzoek was om een pilot haalbaarheidsstudie uit te kunnen voeren van een selectie indicatoren die gebruikt kunnen worden bij het identificeren en kwalificeren van content in een duidingskader.

De geselecteerde content voor deze exercitie werd dus niet gebruikt om inzicht te krijgen in de mate, hoeveelheid of soorten van extremistische en terroristische content op de geselecteerde platforms. Dergelijke vragen werden beantwoord op basis van het literatuuronderzoek. Voor het OSINT onderzoek werden drie platforms geselecteerd - Instagram, TikTok en Reddit - op basis van criteria zoals relevantie voor de Nederlandse context, toegankelijkheid voor open bronnenanalyse, diversiteit van gebruikersdemografie en variatie in moderatiepraktijken (bijv. AI-dominant op TikTok en Instagram, hybride mens-automatische aanpak op Reddit).

Er werd content verzameld met betrekking tot twee nationaal significante gebeurtenissen: de Amsterdamse rellen naar aanleiding van een confrontatie tussen voetbalsupporters van Maccabi Tel Aviv en een groep mensen met sterke sentimenten rond het conflict tussen Israël en Gaza, die bovendien uiting gaven aan antisemitische gevoelens en Maccabi-fans opjaagden en aanvielen (content geplaatst in november 2024) en de "*White Lives Matter*"-projectie op de Erasmusbrug tijdens de Oud-en-nieuwviering (content geplaatst in januari 2023 met betrekking tot het incident van de projectie en december 2024 - januari 2025 met betrekking tot de rechtszaak). Uiteindelijk werden er negen accounts geselecteerd op de drie platforms, die zowel rechts-extremistisch als jihadistisch geïnspireerde narratieven weerspiegelden. Alle berichten die door deze accounts werden gepubliceerd binnen de observatieperiode werden handmatig gescrapet, geanonimiseerd en veilig opgeslagen. Vervolgens werden de berichten gecodeerd met behulp van een pilot codeboek, dat indicatoren testte in de drie categorieën - terroristische, illegale en impliciete extremistische content - wat empirische input leverde over de haalbaarheid van een duidingskader.

Typen Schadelijke Content

Schadelijke online content manifesteert zich in verschillende overlappende categorieën. Voor dit rapport richten we ons op de categorieën die relevant zijn voor terrorisme en radicalisering tot (gewelddadig) extremisme:

- **Terroristische content** wordt gedefinieerd in overeenstemming met Verordening (EU) 2021/784, waarin staat dat terroristische content elk materiaal omvat dat: (i) iemand aanzet of oproept tot het plegen van terroristische strafbare feiten of daaraan bij te dragen, (ii) oproept tot deelname aan activiteiten van een terroristische groepering, (iii) terroristische activiteiten verheerlijkt, onder meer door materiaal te delen dat terroristische aanslagen afbeeldt, of (iv) instructies geeft voor het maken of gebruiken van explosieven, vuurwapens of andere wapens, waaronder chemische, biologische, radiologische of nucleaire stoffen. Daarbij worden terroristische strafbare feiten gedefinieerd overeenkomstig artikel 3 van Richtlijn (EU) 2017/541.

- **Illegale content** verwijst naar online content die in strijd is met nationale of Europese wetgeving. Dit omvat content die op zichzelf illegaal is, evenals content die inbreuk maakt op consumentenbeschermingswetten of een schending vormt van intellectuele eigendomsrechten. Voor de reikwijdte van dit onderzoek richten we ons uitsluitend op

illegale content in de context van terrorisme en gewelddadig extremisme. Dit kan onder meer haatzaaiende uitlatingen omvatten en online content die bijdraagt aan polarisatie en radicalisering.

- **Impliciete extremistische content die schadelijk is:**
  - **'Extremistisch'** verwijst naar content die narratieven verspreidt die groepen of personen buitensluiten of haatdragend van aard zijn, en die kunnen bijdragen aan radicalisering richting terrorisme en (gewelddadig) extremisme.
  - **'Impliciet'** verwijst naar het feit dat de betekenis verhuld is. Wanneer dit opzettelijk gebeurt, is het doel om de illegaliteit, onrechtmatigheid of schadelijkheid van de content te verhullen.
  - **'Schadelijk'** verwijst naar het feit dat de content ernstige schade kan toebrengen aan een individu, een groep mensen, instellingen of de democratische rechtsorde, en dat deze niet wordt beschermd door internationale mensenrechtenrechten.

De categorie impliciete extremistische content is in het bijzonder problematisch. De meerduidigheid van deze content beschermt tegen directe juridische sancties en maakt het eveneens mogelijk verdeeldheid te zaaien en extremistische wereldbeelden te versterken. Bovendien stemmen extremistische actoren hun berichtgeving bewust af om binnen de grijze zone te blijven, wat ervoor zorgt dat content aan moderatie ontsnapt en toch radicaliserend werkt.

Verhulling en Aanpassing

Extremistische actoren gebruiken verhullingsstrategieën om detectie te omzeilen. *Dog whistles*, gecodeerde emoji's en historische of culturele verwijzingen die alleen voor ingewijden begrijpelijk zijn, worden vaak ingezet. Humor en ironie, vooral via memes, dienen zowel als retorisch schild als rekruteringsmiddel. Ze normaliseren extremistische ideeën en weerleggen tegelijk externe kritiek.

Deze actoren zijn wendbaar en adaptief. Naarmate platforms strengere moderatie toepassen op openlijk terroristische content, verschuiven extremisten naar meer impliciete vormen van content waarbij ze zorgvuldig hun discours zo formuleren om het *'awful but lawful'* te laten lijken. Opkomende technologieën versterken deze trend: generatieve AI faciliteert de productie van geavanceerde teksten, beelden en video's, terwijl *deepfake*-technologieën en interactieve game-omgevingen nieuwe mogelijkheden voor verspreiding bieden.

Maatschappelijke Gevolgen

De gevolgen van ongecontroleerde schadelijke content zijn aanzienlijk:

- **Radicalisering**: blootstelling aan extremistische narratieven online is een aantoonbare factor van gewelddadige radicalisering, vooral onder jongeren.
- **Normalisering van geweld**: herhaaldelijke blootstelling aan (gewelddadige) extremistische retoriek vermindert de maatschappelijke weerstand tegen geweld, wat extremistische wereldbeelden inbedt in het reguliere discours.
- **Polarisatie**: schadelijke content verdiept problematische maatschappelijke verdeeldheid, erodeert vertrouwen in instituties en voedt vijandigheid tussen gemeenschappen.
- **Democratische weerbaarheid**: manipulatie van online discours ondermijnt het open democratische debat, wat de ruimte voor pluralisme en constructief meningsverschil vernauwt.

De maatschappelijke uitdaging is dus tweeledig: voorkomen dat onlineplatforms worden misbruikt voor extremistische doeleinden, en tegelijkertijd fundamentele vrijheden, waaronder de vrijheid van meningsuiting, waarborgen.

Het regelgevende en institutionele landschap

Europese Kaders

Als antwoord op deze dreigingen heeft de Europese Unie een gelaagd regelgevend kader ontwikkeld. Twee instrumenten springen eruit:

- **Verordening inzake verspreiding van terroristische online-inhoud (TCO, Verordening EU 2021/784):** verplicht hostingdiensten (HSP's) om terroristische content binnen één uur na melding door bevoegde autoriteiten te verwijderen. De verordening introduceert transparantievereisten, gebruikersmelding en gedifferentieerde verplichtingen afhankelijk van het blootstellingsniveau. Critici wijzen op de onevenredige last voor kleinere platforms die onvoldoende middelen hebben om te voldoen aan de verplichtingen.
- **Digitaledienstenverordening (DSA, Verordening EU 2022/2065)**: verbreedt de regelgevende reikwijdte aanzienlijk om tegen alle illegale content op te treden en legt zeer grote onlineplatforms (VLOP's) en onlinezoekmachines (VLOSE's) verplichtingen op om systemische risicobeoordelingen uit te voeren. De DSA beoogt transparantie te vergroten door jaarverslagen te vereisen, een DSA-transparantieregister te creëren en onderzoekers toegang te geven tot platformgegevens.

Gezamenlijk markeren deze kaders een verschuiving van vrijwillige samenwerking met platforms naar een regelgevingsmodel dat streeft naar harmonisatie, verantwoordelijkheid en minimale waarborgen voor de mensenrechten. Toch blijven er uitdagingen voor de implementatie, met name rond proportionaliteit, capaciteitsbeperkingen van kleinere aanbieders en de bescherming van fundamentele rechten.

Platformbestuur

Naast uitvoering geven aan de juridische verplichtingen, handhaven platforms ook hun eigen gebruiksvoorwaarden (ToS) en communityrichtlijnen. Deze regels reiken vaak verder dan wettelijke verplichtingen en omvatten bredere categorieën van schadelijke content. Hoewel deze proactieve houding kan bijdragen aan het vermijden van bestuurlijke boetes, roept ze zorgen op over het feit dat private actoren in de praktijk de grenzen van online vrijheid van meningsuiting bepalen, zonder duidelijk democratisch toezicht.

Een gebrek aan transparantie verergert dit probleem. Gebruikers begrijpen vaak niet waarom content wordt verwijderd, terwijl onderzoekers en toezichthouders obstakels ondervinden bij het verkrijgen van moderatiegegevens. De ondoorzichtigheid over handhavingsbesluiten vanuit de ToS ondermijnt vertrouwen en verantwoordelijkheid.

Dimensie van Fundamentele Rechten

Het recht op vrijheid van meningsuiting staat centraal in dit debat. Contentmoderatie omvat onvermijdelijk normatieve oordelen over wat toelaatbaar is. Wanneer platforms het zekere voor het onzekere nemen om boetes te voorkomen, lopen zij het risico legitieme kritiek, satire of dissidentie te verwijderen, met onevenredige effecten op gemarginaliseerde of minderheidsgroepen. Daar tegenover staat de ontoereikende moderatie die ertoe kan leiden

dat schadelijke narratieven ongestoord gedijen. Uiteindelijk is het laatste oordeel aan de rechter bij geschillen, maar rechtszaken verlopen traag en zijn gering, en bieden weinig houvast in de tussentijd.

De uitdaging op het gebied van beleid- en regelgeving gaat dus niet alleen over handhaving naleving, maar ook over het inbedden van waarborgen voor fundamentele rechten in moderatiepraktijken.

Onderzoeksvragen

Het WODC heeft, op verzoek van de NCTV, een onderzoek laten uitvoeren naar de vraag of er een betrouwbaar duidingskader kan worden ontwikkeld om platforms te helpen bij het identificeren en modereren van terroristische, illegale en impliciete extremistische content. Er stonden zeven kernvragen centraal in het onderzoek:

1. Welke kenmerken of combinatie van kenmerken in online content bepalen/bepaalt of er sprake is van extremistische of terroristische content op onlineplatforms?
2. Op welke wijze en in welke mate zijn deze kenmerken, al dan niet in combinatie met elkaar, ook te detecteren en te identificeren in de content op onlineplatforms?
3. Wat kan er worden gezegd over de betrouwbaarheid van de detectiemogelijkheden voor extremistische en terroristische content op onlineplatforms?
4. Hoe verhoudt de betrouwbaarheid van de detectiemogelijkheden voor de verschillende typen van schadelijke content op onlineplatforms (terroristisch, (evident) extremistisch en borderline) zich tot het risico op onterechte moderatiebeslissingen door onlineplatforms waarbij het fundamentele recht op vrijheid van meningsuiting wordt geschonden?
5. Is het met de huidige kennis en stand van de techniek mogelijk om een valide en betrouwbaar duidingskader op te stellen voor de detectie en identificatie van zowel de expliciete extremistische en terroristische content als de meer impliciete borderline content op onlineplatforms, zonder dat onterecht het fundamentele recht op de vrijheid van meningsuiting geweld wordt aangedaan?
6. Onder welke condities kan de toepassing van een duidingskader voor extremistische en terroristische online content bijdragen aan de vermindering van zowel de verspreiding van deze schadelijke content als de potentiële radicalisering van internetgebruikers?
7. Indien een valide en betrouwbaar duidingskader niet mogelijk geacht wordt: wat zijn de juridische, technische en eventueel andere knelpunten/barrières die het opstellen van een duidingskader voor extremistische online content in de weg staan? Aan welke voorwaarden zou moeten worden voldaan om deze knelpunten/barrières weg te nemen en wat is de uitvoerbaarheid hiervan?

Deze vragen vormden de analytische lens voor het beoordelen van haalbaarheid, betrouwbaarheid en ethische wenselijkheid.

Belangrijkste Bevindingen

Op basis van het literatuuronderzoek heeft het team essentiële indicatoren om terroristische, illegale en impliciete extremistische online content te kwalificeren geïdentificeerd (onderzoeksvragen 1 en 2), en deze gebruikt om een pilot codeboek te ontwikkelen. Het doel van het onderzoek was echter niet om een volledig duidingskader te ontwikkelen, maar beperkt tot een toetsing van de haalbaarheid ervan. De reikwijdte van het pilot codeboek is daarom bewust beperkt, om - gezien het arbeidsintensieve coderingsproces - deze stap in het onderzoek beheersbaar te houden. De selectie van indicatoren die in het pilot codeboek zijn opgenomen is daarom beperkt tot indicatoren die van belang zijn voor de kwalificatie van rechts-extremistische, jihadistische en impliciete extremistische content. Voor terroristische content gaf het team prioriteit aan typen content die het meest voor discussie vatbaar zijn, terwijl voor illegale content alleen vormen met een mogelijke overlap met impliciete extremistische content werden meegenomen.

Voor de kwalificatie van terroristische content richtte de pilot zich op drie soorten misdrijven: aanzetten tot het plegen of deelnemen aan een terroristisch misdrijf, verheerlijking van terroristische misdrijven en rekrutering voor een terroristische organisatie. De indicatoren zijn ontleend aan de EU-verordening terroristische online-inhoud (EU 2021/784), die bindend is voor zowel platforms als bevoegde autoriteiten binnen de EU. Elk bericht werd gescreend op basis van deze criteria om te beoordelen of het als terroristische content kon worden aangemerkt.

Voor de kwalificatie van illegale content werd slechts een beperkte set niet-terroristische categorieën opgenomen om de haalbaarheid van het duidingskader te testen: aanzetten tot haat, aanzetten tot geweld, en ontkenning, bagatellisering of vergoelijking van internationale misdrijven. Deze werden geselecteerd op basis van hun potentiële overlap met extremistische narratieven en beoordeeld met behulp van indicatoren uit het Nederlandse Wetboek van Strafrecht.

In beide gevallen zorgde een extra waarborg ervoor dat content die onder de vrijheid van meningsuiting valt niet per ongeluk als terroristisch of illegaal werd geclassificeerd. Voor deze laatste stap werd het belangrijk geacht te toetsen of er niet sprake was van satire, een parodie, een artistieke uiting, een gerechtvaardigde bijdrage aan het publieke debat, of een legitieme referentie naar een herdenking van een historisch feit, het koloniale verleden, of dekolonisatie.

Indicatoren van Impliciete Extremistische Content

Waar het onderzoeksteam voor de vorige categorieën kon verwijzen naar wettelijke kaders om essentiële indicatoren te identificeren, was dit niet het geval voor impliciete extremistische content. Het onderzoek heeft daarom, op basis van een analyse van literatuur en beleidsrapporten, verschillende relevante indicatoren voor het classificeren van impliciete extremistische content geïdentificeerd. De voorgestelde categorieën van indicatoren om impliciete extremistische content te identificeren zijn:

- **Verhulling van Betekenis (CM)**: opzettelijke verhulling via ironie, humor of codetaal.
- **Schadelijke Allianties/Affiliaties (HA)**: verwijzingen naar extremistische groepen, ideologieën of symbolen.
- **Problematische Verwijzingen (PR)**: het aanhalen van historische of actuele gebeurtenissen met een extremistische framing.
- **Impliciete Actietriggers (AT)**: subtiele signalen die het publiek aansporen om actie te ondernemen.
- **Verondersteld Oogmerk Schade te berokkenen (IH)**: verondersteld oogmerk schade te berokkenen met de geplaatste content.

Elke categorie bestond uit meerdere verschillende indicatoren. Om content te kwalificeren als impliciete extremistische content, moest er een combinatie van indicatoren aanwezig zijn. Ook hier werden aanvullende waarborgvragen ingebouwd, die toetsen of er sprake was van satire, een parodie, een artistieke iting, een gerechtvaardigde bijdrage aan het publieke debat, of een legitieme referentie naar een herdenking van een historisch feit, het koloniale verleden, of dekolonisatie, om er op deze wijze voor te zorgen dat content die onder de vrijheid van meningsuiting valt, werd uitgezonderd.

De indicatoren van terroristische, illegale en impliciete extremistische content werden opgenomen in een pilot codeboek om hun betrouwbaarheid te testen (onderzoeksvragen 3 en 4). De indicatoren werden beoordeeld op drie criteria: bruikbaarheid, nuttigheid en interpreteerbaarheid.

Bruikbaarheid (*usability*) weerspiegelt hoe eenvoudig en tijdsefficiënt het is om een indicator te coderen. Groen betekent gemakkelijk, oranje tot op zekere hoogte veeleisend en rood wordt gebruikt voor lastige indicatoren. *Nuttigheid* (*usefulness*) geeft aan of een indicator een betekenisvolle bijdrage levert aan de algehele beoordeling. Rood duidt op overbodigheid en oranje geeft aan dat er beperkte bijdrage of overlap is. *Interpreteerbaarheid* (*interpretability*) evalueert de mate van subjectiviteit van de codering. Groen staat voor duidelijke indicatoren, oranje voor indicatoren die tot op zekere hoogte subjectief zijn en verfijning vereisen, en rood voor zeer subjectieve indicatoren, geïdentificeerd door discrepanties tussen codeurs.

Het onderzoek reflecteert vervolgens op de resultaten van deze evaluatie die de volgende bevindingen opleverde:

Het Belang van Context

Indicatoren functioneren zelden geïsoleerd. Interpretatie hangt sterk af van de context, waaronder de identiteit van de gebruiker, culturele referenties en hoe de content door gebruikers wordt ervaren. Deze contextafhankelijkheid compliceert opschaling van detectiemechanismen en ondermijnt de betrouwbaarheid wanneer er uitsluitend gebruik wordt gemaakt van geautomatiseerde systemen.

Hybride Detectiemodellen

Het onderzoek onderstreept dat geautomatiseerde hulpmiddelen de nuance van impliciete extremistische content niet kunnen herkennen. Een hybride model - een combinatie van AI-ondersteunde voorselectie en menselijke expertise - is essentieel. Menselijke codeurs brengen contextgevoeligheid, maar hebben gestructureerde begeleiding, training en waarborgen tegen bias nodig.

Operationele Keuzes van de Technologiesector

Platforms vertrouwen in toenemende mate op geautomatiseerde hulpmiddelen als ruggengraat van hun moderatiesystemen en presenteren dit als oplossing voor schaal- en efficiëntieproblemen. Uit het onderzoek blijkt echter dat, hoewel automatisering nuttig is voor het detecteren van openlijk terroristische of illegale content, het slecht presteert bij toepassing op impliciete extremistische content dat vaak gebaseerd is op culturele nuances, ironie of gecodeerde referenties. De afname van menselijke moderators bij grote platforms verergert dit risico, waardoor er een groeiende kloof ontstaat tussen de complexiteit van schadelijke content en de methodes die de sector heeft gekozen om dit aan te pakken.

Transparantie en Verantwoording

Het gebrek aan transparantie van platforms over moderatiepraktijken beperkt het publieke vertrouwen en de academische controle ernstig. Er zijn diverse moderatiemogelijkheden voor zowel content als accounts, waaronder het beperken van de zichtbaarheid. Iedere optie heeft een andere invloed op de vrijheid van meningsuiting. Het blijft echter onduidelijk hoe vaak welke moderatiebeslissingen worden genomen en in welke mate rekening wordt gehouden met de impact die deze beslissingen hebben op de vrijheid van meningsuiting (proportionaliteitsvereiste). Zonder duidelijke rapportage en toegankelijke beroepsmogelijkheden blijven gebruikers in het ongewisse over de regels die hun online uitingen beheersen.

Haalbaarheid van een Duidingskader?

Op basis van de bevindingen zijn er verschillende fundamentele uitdagingen geïdentificeerd. Gezamenlijk wijzen deze uitdagingen erop dat, hoewel stapsgewijze verbeteringen mogelijk zijn, een universeel en volledig betrouwbaar duidingskader op dit moment onhaalbaar is.

Uitdagingen

Met betrekking tot de haalbaarheid van een betrouwbaar duidingskader werden de volgende uitdagingen geïdentificeerd:

- **Meerduidige Definities en Vage Grenzen**
  - Gebrek aan universeel aanvaarde wettelijke definities van termen als ‘terrorisme’, ‘aanzetten tot haat’, ‘gewelddadig extremisme’ of ‘opruiing’.
  - Gebrek aan internationaal erkende definities voor essentiële begrippen zoals ‘groep’, ‘legitiem’ of ‘zelfverdediging’.
  - Onduidelijke referentiekaders voor wanneer content als schadelijk versus aanvaardbaar wordt beschouwd.
  - Lastig om onderscheid te maken tussen extremistische retoriek en satire, politieke kritiek of legitiem debat.
  - Vage categorieën ondermijnen consistentie in codering en overeenkomst tussen codeurs.
  - Het impliciete karakter van content kan betrekking hebben op het verhullen van de schadelijkheid, rechtmatigheid of illegaliteit van de content. Begrijpen wanneer dit opzettelijk gebeurt is lastig.

- **Veranderende Verhullingstactieken**
  - Extremistische actoren passen taalgebruik en strategieën snel aan, waardoor statische duidingskader achterhaald raken.
  - *Dog whistles*, ironie, memes en gecodeerde termen ontlopen detectie.

- **Risico Subjectiviteit en Vooringenomenheid bij Menselijke Moderatoren**
  - De sterke afhankelijkheid op persoonlijke interpretatie en context leidt tot inconsistente resultaten.
  - Risico dat gewone uitingen van ongenoegen of oppositie abusievelijk als extremistisch worden geclassificeerd.
  - De culturele, ideologische of politieke achtergronden van codeurs kunnen hun oordelen beïnvloeden.
  - Indicatoren zijn kwetsbaar voor misbruik door bevooroordeelde codeurs of *flaggers*.

- **Nauwkeurigheid v. Vooringenomenheid**
  - Geautomatiseerde moderatie implementeert de vastgestelde algoritmes weliswaar nauwkeuriger, maar het risico bestaat dat er sprake is van ingebouwde vooringenomenheid die zich niet snel manifesteert.
  - Geautomatiseerde moderatie is weliswaar goedkoper en sneller, maar vaak minder goed in staat om impliciete extremistische content accuraat te detecteren, daarentegen is menselijke beoordeling beter geschikt voor de beoordeling van impliciete extremistische content, maar is het tevens duurder en brengt het risico van vooringenomenheid met zich mee.
  - Het enorme volume aan online content overstijgt de capaciteit van zowel menselijke codeurs als huidige AI-systemen.

- **Contextafhankelijkheid en Arbeidsintensiteit**
  - Veel indicatoren vereisen diepgaande kennis van gebruikersgeschiedenis, ideologie of platformdynamiek. Dit beperkt de bruikbaarheid en maakt identificatie lastig.

Een bijkomende structurele uitdaging zijn de onderliggende bedrijfsmodellen die de praktijken van hostingdiensten (HSP's) vormgeven. Grote technologiebedrijven zijn primair winstgedreven en hebben weinig behoefte om te investeren in menselijke moderatoren die de betrouwbaarheid zouden verbeteren en fundamentele rechten zouden kunnen waarborgen. Dit onderzoeksteam stuitte, net als vele anderen, op aanzienlijke terughoudendheid van platforms om mee te werken of transparantie te bieden over hun moderatieaanpak. Dit onderstreept de verantwoordingskloof tussen private zelfregulering en het publieke belang. De inkrimping van menselijke moderatieteams, gecombineerd met de ondoorzichtigheid van algoritmische besluitvorming, ondermijnt niet alleen de betrouwbaarheid van detectie, maar beperkt ook de mogelijkheden voor democratisch toezicht.

Aanbevelingen

Een universeel geldig en betrouwbaar duidingskader voor het detecteren van impliciete extremistische content lijkt niet haalbaar (onderzoeksvragen 5-7). Meer betrouwbare en adaptieve kaders zouden echter wel haalbaar kunnen zijn voor gebruik door hostingdiensten (HSP's). Door definities te verfijnen, complexe indicatoren te operationaliseren, en iteratief leren, hybride systemen en gezamenlijk toezicht in te bedden, kunnen de gebruikte indicatoren worden omgevormd tot consistentere, nuttigere en betrouwbaardere hulpmiddelen. Een dergelijke aanpak balanceert de noodzaak om impliciete extremistische content te identificeren die de vrijheid van meningsuiting waarborgt. Uiteindelijk, en zolang er geen aanvullende wettelijke kaders van toepassing zijn, zal de effectiviteit van elk detectiemechanisme afhangen van de bereidheid en capaciteit van HSP's om het verantwoord toe te passen. Overheden zouden, in samenwerking met de EU, hun dialoog met HSP's moeten intensiveren om dit proces te stimuleren.

Of een dergelijk universeel duidingskader überhaupt wenselijk is, blijft een open vraag. Ethische overwegingen moeten een centrale rol spelen bij die afweging. Daarbij speelt het feit dat onze huidige vorm van communicatie steeds complexer is geworden: online- en offline-werelden zijn nauw verbonden, en zijn een weerspiegeling van zowel sociale diversiteit als groeiende polarisatie. Deze snelle transformatie heeft het publieke debat over welke normen en waarden gelden op sociale mediaplatforms ingehaald. Gezien het feit dat wettelijke kaders ambigu of inconsistent zijn in het stellen van grenzen, zou de ontwikkeling van ieder duidingskader voor online content moeten beginnen met een brede maatschappelijke dialoog over wat acceptabele uitingsvormen zijn en wanneer dit overgaat naar onacceptabel en schadelijk uitingsvormen.

Ondanks deze bezwaren benadrukken we hieronder aanbevelingen voor de technologiesector en voor beleidsmakers. Deze aanbevelingen volgen voor het merendeel rechtstreeks uit de bevindingen van de studie. Echter, de expertise en praktische ervaring die de onderzoekers hebben opgedaan bij het implementeren van preventieprogramma's en bij capaciteitsopbouw, hebben ook bijgedragen aan de formulering van enkele aanbevelingen.

Basisvoorwaarden die HSP's kunnen implementeren voor een betrouwbaar en accuraat duidingskader:

1. **Duidelijke Definities en Referentiekaders**
   a. Leef de verplichting na om precieze definities voor vage termen te hanteren;
   b. Stel referentiekaders vast voor opruiing, haat of vijandigheid, die niet in strijd zijn met de vrijheid van meningsuiting;
   c. Wees transparant over de combinatie van indicatoren die aanwezig moeten zijn voor de vaststelling van impliciete extremistische content.

2. **Richtlijnen en Voorbeelden**
   a. Geef illustratieve voorbeelden van wat wel en niet geoorloofde content is ten aanzien van het gehele ideologische spectrum, inclusief satire, geoorloofde kritiek en schadelijke retoriek.
   b. Gebruik beslisbomen of coderingsstroomdiagrammen om toepassing te standaardiseren.
   c. Ontwikkel typologieën van wat of wie tot de *in- or out-group* behoort om accurate toepassing van de betreffende indicator te vergemakkelijken.

3. **Beperking van Subjectiviteit en Vooringenomenheid**
   a. Voorzie codeurs van training, prompts en strategieën om vooringenomenheid tegen te gaan.
   b. Pas beoordelingsprotocollen met het "vierogenprincipe" en consensusmethoden toe voor de categorie content die gekwalificeerd kan worden als impliciete extremistische content.

4. **Operationalisering van Complexe Indicatoren**
   a. Splits brede indicatoren op per ideologische stroming en in subcategorieën.
   b. Gebruik triangulatie met contextuele aanwijzingen (bijv. geschiedenis, platformdynamiek) om betrouwbaarheid te verbeteren.

5. **Iteratieve en Adaptieve Kaders**
   a. Gebruik voor de verschillende ideologische stromingen aparte duidingskaders, in plaats van een universeel duidingskader.
   b. Beschouw duidingskaders als "levende documenten" die mee ontwikkelen met nieuwe extremistische tactieken.
   c. Update de duidingskaders regelmatig aan de hand van nieuwe inzichten en voorbeelden uit onderzoek of naar aanleiding van resultaten die uit evaluaties komen.

6. **Hybride AI-Menselijke Systemen**
   a. Combineer AI-voorselectie voor schaalbaarheid met deskundige menselijke beoordeling voor contextgevoelige beoordelingen.
   b. Gebruik LLM's en databanken om arbeidsintensieve taken uit te voeren, maar houd mensen ondertussen betrokken bij het proces.

7. **Multistakeholder-Samenwerking**
   a. Werk regelmatig samen met onderzoekers, praktijkdeskundigen, platforms, het

maatschappelijk middenveld en gemarginaliseerde groepen bij de ontwikkeling en herziening van duidingskaders.
   b. Bouw een brede consensus om bias, overschrijding en misclassificatie te beperken.

8. **Transparantie en Beroepsprocedures**
   a. Verbeter transparantie en rapportage over moderatiebeslissingen.
   b. Respecteer het proportionaliteitsbeginsel bij moderatiebeslissingen om de vrijheid van meningsuiting beter te waarborgen: Om de vrijheid van meningsuiting te waarborgen, is het nodig om een breder scala aan proportionele moderatieopties te ontwikkelen en ook daadwerkelijk toe te passen in de praktijk.
   c. Geef heldere informatie over hoe en waar men in beroep kan gaan tegen een moderatiebesluit.

Specifieke Aanbevelingen voor de Beleidsmakers

Dit onderzoek is uitgevoerd op verzoek van de NCTV. Hoewel de bevindingen en aanbevelingen van dit onderzoek relevant zijn voor een breder publiek, speelt de NCTV, als coördinerende actor, een sleutelrol in het vormgeven en implementeren van beleid in Nederland, in de samenwerking met Europese partners om het Europese beleid te bevorderen en in de dialoog met HSP's. Op basis van de bevindingen van dit onderzoek en gebaseerd op de expertise van de leden van het onderzoeksteam, formuleren wij daarom aanbevelingen voor beleidsmakers die in het bijzonder een coördinerende rol vervullen, zoals de NCTV en de ATKM.

Beleidsmakers wordt aanbevolen om:

1. **Af te zien van de term 'borderline' content**, omdat dit de verwarring over de reikwijdte en betekenis van de term alleen maar zal vergroten.

2. **Een *multistakeholdergroep* op te richten**, bestaande uit onderzoekers, praktijkexperts, platforms, het maatschappelijk middenveld en gemarginaliseerde groepen om (regelmatig) en op een transparante wijze te reflecteren op en publiekelijk te rapporteren over:
   a. Een reeks indicatoren om terroristische, illegale en impliciete extremistische content te detecteren, in lijn met het idee van het bijhouden van een levend document;
   b. De indicatoren die volgens dit onderzoek als problematische worden beschouwd en hun formulering te verbeteren;
   c. De kritische grens voor de combinatie van indicatoren die nodig zijn voor de kwalificatie als impliciete extremistische content
   d. Kenmerkende gebeurtenissen of historische feiten, specifiek voor de Nederlandse context, alsmede veelgebruikte uitdrukkingen, specifiek taalgebruik, en codetaal dat eigen is aan de Nederlandse taal en gehanteerd wordt door extremistische groepen die online content in het Nederlands plaatsen, wat van belang kan zijn voor de contextuele interpretatie van impliciete extremistische content.

3. **Een publiek debat te bevorderen**
   a. Over wat schadelijke en onrechtmatige content is en wat schadelijke maar rechtmatige content is;
   b. Over hoeveel autonomie HSP's hebben bij het beheren van onze openbare ruimte voor publieke meningsuiting.

4. **Media-geletterdheid programma's op scholen, preventieprogramma's voor jongerenorganisaties en strategische communicatieve de-escalatietechnieken te ondersteunen**, die inzetten op het tegengaan van verspreiding van dergelijke schadelijke content, en het vergroten van weerbaarheid, in het bijzonder voor minderheids- en gemarginaliseerde groepen.

5. **Een helder strategisch communicatiebeleid te ontwikkelen** om adequaat te reageren bij grootschalige verspreiding van schadelijke maar rechtmatige content, inclusief uitleg waarom iets als schadelijk wordt beschouwd. Tevens **publiekelijk stelling te nemen tegen schadelijke onrechtmatige content die gericht is op gemarginaliseerde groepen.** Daarnaast is het van belang lokale beleidsmakers hierin te ondersteunen.

De coördinerende overheidsactoren (zoals de NCTV of de ATKM) worden aanbevolen om in dialoog met HSP's

6. **Op transparante wijze samen te (blijven) werken met grote en kleine hostingdiensten:**
   a. Om een open discussie te voeren over de indicatoren die zij gebruiken in hun duidingskaders en of zij verschillende duidingskaders gebruiken voor verschillende ideologieën;
   b. Om informatie-uitwisseling en transparantie over de toepassing van verschillende proportionele moderatiebeslissingen en de impact op de vrijheid van meningsuiting te bevorderen.

7. Zonder de HSP's van hun primaire verantwoordelijkheid te ontdoen, maar rekening houdend met het feit dat het Nederlands een kleine taal is**, een lijst met veelgebruikte uitdrukkingen, specifiek taalgebruik, codetaal die specifiek zijn voor het Nederlands** en worden gebruikt door extremistische groeperingen die online actief zijn in Nederland - zoals besproken in de multistakeholdergroep - **te delen,** om zo te helpen bij de contextuele interpretatie van impliciete extremistische content.

8. **Regelmatig contextuele achtergrondinformatie te verstrekken** aan HSP's over typisch Nederlandse (actuele of historische) gebeurtenissen, die hen kunnen helpen bij de contextuele interpretatie van online content.

De coördinerende overheidsactoren wordt aanbevolen, om in dialoog met andere Europese lidstaten en de Europese Commissie:

9. **Samen met de sector een sectorbrede gedragscode te ontwikkelen,** die een keurmerk aanbiedt dat consumenten een beter inzicht biedt in hoe goed HSP's scoren op detectie en moderatie; die normen vaststelt voor het percentage van menselijke beoordelingen en moderatiebeslissingen; en die duidelijkheid schept over de termen die in de ToS gebruikt worden, en ten aanzien van de transparantie van moderatiebeslissingen en de beroepsprocedures.

10. **Zich in te zetten voor verdere versterking van Europese regelgevingskaders** door meer transparantie en verantwoordingsplicht van HSP's te eisen, een ex-ante evaluatie te eisen over hoe HSP's de vrijheid van meningsuiting respecteren door hun ToS toe te passen, regelmatige ex-post evaluaties te eisen over hoe die vrijheid is gerespecteerd in moderatiebeslissingen, strengere regels te hanteren voor moderatiemethodes (AI versus handmatig), en door heldere definities en richtlijnen te bieden.

Conclusie

Het onderzoek concludeert dat de ontwikkeling van een universeel, betrouwbaar duidingskader voor de detectie van terroristische, illegale en impliciete extremistische online content niet haalbaar is gezien de huidige juridische, technische en ethische omstandigheden. Meerduidigheid van definities, contextuele complexiteit en de steeds veranderende tactieken van extremistische actoren ondermijnen de betrouwbaarheid van dergelijke kaders en vergroten de risico's voor de vrijheid van meningsuiting.

Toch betekent de onhaalbaarheid van een universeel duidingskaders niet dat vooruitgang onmogelijk is. Stapsgewijze verbeteringen zijn mogelijk door duidelijkere definities, adaptieve indicatoren, hybride AI-menselijke modellen en samenwerking tussen *multistakeholders'*. In plaats van te zoeken naar een pasklare oplossing, zouden platforms en beleidsmakers flexibele, dynamische aanpakken moeten nastreven die afgestemd zijn op specifieke contexten en ideologieën.

Voor de NCTV en de bredere beleidswereld reikt de uitdaging verder dan alleen detectie, maar raakt het ook aan bredere maatschappelijke vraagstukken. Wat wordt beschouwd als schadelijke maar rechtmatige (*awful but lawful*) content? Hoeveel autonomie mogen private bedrijven hebben in het besturen van online publieke ruimtes? Hoe kunnen samenlevingen een balans vinden tussen veiligheid en vrijheid?

De weg voorwaarts vereist voortdurende dialoog, transparantie en adaptiviteit. Naarmate extremistische actoren blijven innoveren, moeten beleidsmakers, platforms en gemeenschappen dat ook doen. Weerbaarheid opbouwen vereist niet alleen technologische oplossingen, maar ook een democratisch debat over de normen van onlinecommunicatie en de waarden die samenlevingen willen handhaven.

1. Introduction

1.1 Background

In the Terrorist Threat Assessment Netherlands (DTN) of June 2024,[1] a warning is issued about terrorists and extremists who use social media and other online platforms to share terrorist and extremist content in order to spread and reinforce their ideology. In addition to terrorist and extremist content, a category of content is distributed that is oftentimes referred to as 'borderline' content. This content is generally considered harmful because of its effect on a specific group or because of its undermining impact on democracy, yet does not - at first sight- fall under the legal definitions of terrorist content or illegal content, and therefore is considered to fall under the protection of freedom of expression.

A complicating factor, related to the spread of online terrorist, extremist or borderline content, is that concealing tactics are used to avoid detection of the concealed unlawful elements in the content and, henceforth, the moderation of the content. Furthermore, wherever possible, terrorists and extremists make use of available technological developments to create and spread the content. The online platforms are used as means of communication between terrorists, as well as for financing, training, and planning attacks, recruitment, and propaganda purposes.

With the emergence and development of generative artificial intelligence (AI) and large language models (LLMs), harmful content is being used by terrorist or extremist groups or individuals in even greater volumes, more rapidly, more precisely targeted, and in multiple languages. Jihadist groups such as Islamic State (ISIS) and Al-Qaeda are already using AI for propaganda purposes, [2] and right-wing extremists (RWX) in Western countries are increasingly making use of generative AI as well. [3]

The dissemination of terrorist, extremist, and borderline content can incite hatred, normalise terrorist violence, and contribute to societal polarisation. To gain control over harmful content, it is crucial to intervene as early as possible in the dissemination process, yet ensure that the freedom of expression is protected. Intervention can take the form of preventive measures, content moderation, or even criminal prosecution where necessary.

There are various forms of harmful content. Some are easier to recognise as terrorists or illegal, while others are more ambiguous or even purposefully concealed. The category of content that does not at first glance belong to terrorist or illegal content, but is still considered harmful, is often referred to as borderline content. The harmfulness of the content can furthermore be either explicit or implicit. The latter is often disguised in humour, irony, dog whistles, or memes in various forms. The fact that this borderline content is initially not recognised as falling within the categories of terrorist or illegal content, does not definitely preclude its illegality or unlawfulness. It merely merits a more thorough inspection and understanding.

To address this, policies have been developed that combat terrorist and illegal online content that is deemed harmful. At the European level, the Terrorist Content Online Regulation (TCO) and the Digital Services Act (DSA) are in force. These regulations impose various obligations on online platforms and search engines to take measures in handling this type of harmful content.

1 "Dreigingsbeeld Terrorisme Nederland," *NCTV*, published June 17, 2025, https://www.nctv.nl/onderwerpen/dtn.
2 Daniel Siegel, "AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's Generative AI Digital Arsenal," *GNET Insights* (2024), https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/.
3 Bàrbara Molas, and Heron Lopes, "Say it's only fictional: How the far-right is jailbreaking AI online and what can be done about it," *ICCT* (October 2024), https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it.

The TCO also gives national authorities—such as the Dutch Authority for the prevention of online Terrorist Content and Child Sexual Abuse Material (ATKM)—the power to order platforms to remove content within one hour.

Additionally, there is European-level cooperation through the EU Internet Forum and EUROPOL's Internet Referral Unit, using the PERCI system. However, to potentially avoid the heavy fines that national authorities can impose, many online platforms already proactively remove large volumes of content that violate their own Terms of Service (ToS). To be clear, the ToS of the various platforms provide their own interpretation of existing binding legal frameworks, as well as their own interpretation of content that they otherwise consider to be harmful. This may thus include different definitions of terrorist, illegal, extremist, hateful or borderline content.

The WODC (Research and Data Centre, a Dutch agency in the field of Justice and Security), at the request of the Dutch National Coordinator for Counterterrorism and Security (NCTV), has commissioned this study into the feasibility of the development of an assessment framework of online harmful content – both explicit and implicit – that can contribute to radicalisation towards extremism and terrorism. If considered feasible, such an assessment framework could assist HSPs to execute more accurate detection of such content, and to take moderation decisions without curtailing the freedom of expression.

1.2 Setting the Scene

National authorities, regional bodies, and civil society are increasingly concerned about the relatively easy access to all forms of harmful content online and the growing normalisation of extremist ideologies on the internet.[4] Exposure to such content amplifies radical attitudes and increases risks of political violence, particularly within right-wing extremist and radical Islamist movements.[5] Offline push factors also play a significant role in radicalisation, making extreme online messages more impactful when these real-world factors are present.[6] Additionally, individuals exposed to extremist ideologies online may engage in real-world events or interactions that further reinforce these messages, highlighting the complex interaction between online and offline influences.[7]

1.2.1 The Proliferation of Harmful Online Content

Two prominent events have raised concerns of authorities regarding the spread of terrorist and extremist content online: the online announcement and live streaming of the shootings in the Christchurch mosques and in Halle,[8] both in 2019. The European Commission reports an 18 percent increase in terrorist and violent extremist content and a 65 percent increase in borderline content[9] on social media between July 2022 and May 2023.[10]

4 See e.g. Europol, *European Union Terrorism Situation and Trend Report 2024 (EU TE-SAT)* (Luxembourg: Publications Office of the EU, 2025), https://www.europol.europa.eu/publication-events/main-reports/european-union-terrorism-situation-and-trend-report-2024-eu-te-sat; Ron van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms? Ontwikkelingspaden en handelingsperspectieven* (Utrecht: Verwey Jonker Instituut, 2023), https://repository.wodc.nl/bitstream/handle/20.500.12832/3304/3341-rechtsextremisme-op-sociale-media-platforms-volledige-tekst.pdf?sequence=1&isAllowed=y.
5 Van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms?* 42.
6 Van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms?* 36.
7 William Allchorn, "Turning Back to Biologised Racism: A Content Analysis of Patriotic Alternative UK's Online Discourse," *GNET Insights* (2021), https://gnet-research.org/2021/02/22/turning-back-to-biologised-racism-a-content-analysis-of-patriotic-alternative-uks-online-discourse/.
8 Daniel Koehler, "The Halle, Germany, Synagogue Attack and the Evolution of the Far-Right Terror Threat," *CTC Sentinel* 12, no. 11 (December 2019), https://ctc.westpoint.edu/halle-germany-synagogue-attack-evolution-far-right-terror-threat/.
9 According to the EU Internet Forum Handbook on Borderline Content (2023) 'borderline content' is defined as "content that is not explicitly promoting or supporting terrorism and violent extremism, but which may contain language or ideas that could be leading towards pathways of radicalisation. The criteria for this category include: content that is hard to identify as illegal or as related to violent extremism and radicalisation; content that, despite being legal, can harm and lead to violent extremist behaviour and radicalisation (such as disinformation, conspiracy theories, which can also lead towards dehumanisation); and tactics used to manipulate borderline content leading to violent extremism, such as algorithmic amplification techniques that profit from biases in content sharing algorithms."
10 Anna Maria Carpani et al., *EU Internet Forum: Study on the Role and Effects of the Use of Algorithmic Amplification to Spread Terrorist, Violent Extremist and Borderline Content* (Luxembourg: Publications Office of the EU, 2023), 2, https://data.europa.eu/doi/10.2837/259157.

Researchers have highlighted the relationship between exposure to online extremist content and violent radicalisation and found that the internet and social media play a significant role in extremist violence.[11] Young individuals are particularly vulnerable to online recruitment by extremist and terrorist groups. A recent survey by the European Parliament found that young individuals primarily obtain information through social media platforms, which increases concerns about their susceptibility to online radicalisation.[12]

Terrorist and extremist groups have adapted their methods using new technologies such as artificial intelligence, deepfakes, or popular video games to disseminate extremist narratives.[13] (In chapter 5, the scope of the terrorist, extremist and borderline content will be further elaborated upon.) Such content is being shared on social media, streaming and gaming platforms,[14] or file-sharing websites, all of which are openly accessible and shared in various forms and contexts.[15] It can range from text-based messages and posts, image-based communication, including memes, to video clips and links to external websites. Notably, this content has moved from largely inaccessible sites a decade ago to openly accessible websites.[16]

While some content is apparently terrorist or extremist in nature, for example, the display of symbols of designated terrorist organisations or calls to ideologically motivated violence, other content is less explicit, although not less dangerous. Research has found that extremists of all different ideologies often rely on coded language, symbols, and humour, concealing the true meaning of their messages.[17] Hence, a lot of content spread by terrorists and extremists online at first glance seems 'awful but lawful', also referred to as borderline content.[18] Notably, research indicates that with increasing moderation of evidently terrorist online content, extremists tend to alter the tone of their messaging, deliberately moving to spread borderline content to avoid content moderation.[19] Therefore, if online service providers want to contribute in a responsible manner to keeping the internet a safe space, they must adjust their content moderation tools to effectively address borderline content, including instances where users employ ambiguous or coded language. However, the efficacy of content moderation is frequently compromised by a lack of expertise on the topic and the complexity of determining what constitutes borderline content.[20]

1.2.2 Evolving Regulatory Frameworks

In response to the rising threat of harmful content online described above, certain national governments and the European Union have developed a layered regulatory framework with the aim of limiting the dissemination of such content in the online sphere.

11 Ghayda Hassan et al., "Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence," *International Journal of Developmental Science* 12, no. 1-2 (2018): 71, https://doi.org/10.3233/DEV-170233.
12 European Parliament, and Ipsos European Public Affairs, "Youth Survey 2024," *Flash Eurobarometer* (February 2025): 4, https://europa.eu/eurobarometer/surveys/detail/3392.
13 Europol, *EU TE-SAT 2024*, 6.
14 "Steam-Powered Hate: Top Gaming Site Rife with Extremism & Antisemitism," *ADL Center on Extremism* (November 2024), https://www.adl.org/resources/report/steam-powered-hate-top-gaming-site-rife-extremism-antisemitism.
15 Tech Against Terrorism, "Patterns of Terrorist Online Exploitation," *TCAP Insights* (April 2023): 26, https://26492205.fs1.hubspotusercontent-eu1.net/hubfs/26492205/260423%20TCAP%20INSIGHTS%20-%20FINAL.pdf.
16 For an assessment of online threats posed by terrorist in the early 2010s, see e.g. Bibi van Ginkel, "Responding to Cyber Jihad: Towards an Effective Counter Narrative," *ICCT* (March 2015), https://www.icct.nl/sites/default/files/2022-12/ICCT-van-Ginkel-Responding-To-Cyber-Jihad-Towards-An-Effective-Counte.pdf.
17 Bàrbara Molas, "Alt-solutism: Intersections between Alt-Right Memes and Monarchism on Reddit," *ICCT* (February 2023), https://www.icct.nl/publication/alt-solutism-intersections-between-alt-right-memes-and-monarchism-reddit; Maik Fielitz, and Reem Ahmed, "It's not funny anymore. Far-right extremists' use of humour," *Radicalisation Awareness Network* (2021), https://home-affairs.ec.europa.eu/system/files/2021-03/ran_ad-hoc_pap_fre_humor_20210215_en.pdf.
18 Bibi van Ginkel et al., "Online Monitoring of Radicalisation and (Violent) Extremism: Mapping Legal and Policy Challenges for Online P/CVE Work," *RAN Conclusion Paper* (2023). This paper is not publicly available, but on file with the research team.
19 Ye Bin Won, and Jonathan Lewis, "Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts," *GNET Insights* (2021), https://gnet-research.org/2021/04/06/male-supremacism-borderline-content-and-gaps-in-existing-moderation-efforts/; Allchorn, "Turning Back to Biologised Racism."
20 Van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms?* 16.

Notably, the German Network Enforcement Act (NetzDG)[21] is considered to be one of the first laws globally to mandate social media platforms to remove illegal content under strict timelines, imposing harsh financial penalties in case of non-compliance.[22] However, the NetzDG has also been criticised for being a blueprint for authoritarian regimes to suppress democratic dissent[23] as well as its risk to lead to so-called overblocking,[24] meaning the removal of content that is permissible.

At the EU-level, the 2021 Regulation on Dissemination of Terrorist Content Online (TCO, Regulation EU 2021/784)[25] follows a similar approach to the NetzDG concerning fines in case hosting service providers (HSPs) do not comply with the one-hour deadline of a removal order issued under Article 3 TCO. In doing so, the TCO applies to all HSPs offering services in the EU, irrespective of their physical presence, where a "substantial connection" to the Member State exists.[26] It only relates to terrorist content encompassing, among others, incitement, glorification, recruitment, and preparatory acts to commit a terrorist act.[27] The TCO further imposes obligations on HSPs regarding content preservation following a removal order, transparency through annual reporting, user notification, and the establishment of accessible appeal mechanisms. It also introduces a differentiated regulatory burden by triggering further compliance obligations once an HSP has been identified as being exposed to terrorist content by a competent national authority. Nevertheless, the TCO has been criticised for placing an unequal burden on small and medium-sized HSPs that often lack resources to comply with the strict removal order obligations or to carry out content moderation at their own initiative, as to why the European Commission initiated several projects supporting small and medium-sized HSPs.[28]

The Digital Services Act (DSA, Regulation EU 2022/2065)[29] builds on the TCO's architecture but significantly broadens both the scope of content and the depth of regulatory mechanisms. It introduces a layered framework based on service size and function, distinguishing between HSPs, online platforms, and very large online platforms (VLOPs) and very large search engines (VLOSEs) with more than 45 million monthly users in the EU.[30] The DSA goes beyond terrorist content to regulate all forms of illegal content, including hate speech, defamation, and consumer fraud, both pursuant to Union and national laws. It obligates platforms to act on content flagged by trusted flaggers or public authorities and to clearly define moderation policies in their ToS. Importantly, the DSA builds on the TCO's transparency provisions by obliging service providers to issue annual transparency reports on their content moderation efforts, allowing researchers

21 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), I no. 61 Bundesgezets (2017), https://www.bgbl.de/xaver/bgbl/start.xav#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl117s3352.pdf%27%5D__1736937938042.
22 Sabrina Maaß et al., "Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook," *Telecommunications Policy* 48, no. 5 (June 2024): 4, https://doi.org/10.1016/j.telpol.2024.102719.
23 Jacob Mchangama, and Joelle Fiss, "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship," *Justitia* (November 2019): 6, https://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf; For a more recent, extensive list of countries adopting regulations based on the NetzDG, see Jacob Mchangama, and Natalie Alkiviadou, "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two," *Justitia* (September 2020), https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf.
24 Maaß et al., "Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook."
25 On addressing the dissemination of terrorist content online (TCO), regulation (EU) 2021/784 European Parliament and Council of the EU (2021), https://eur-lex.europa.eu/eli/reg/2021/784/oj.
26 Pursuant to Art. 2 (5) TCO, such a substantial connection can arise either from "having a significant number of users [...] in one or more Member States" or from the fact that the provider is "targeting [...] its activities to one or more Member States". Other indicators for whether the service is targeted at one or more MS according to the preamble, could be "the availability of an application in the relevant national application store, [...] local advertising or advertising in the language used in that Member State, or [...] the handling of customer relations such as by providing customer service in the language generally used in that Member State."
27 In defining terrorist content, the TCO relies on existing EU legislation on terrorism. On combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, directive (EU) 2017/541 European Parliament and Council of the EU (2017), https://eur-lex.europa.eu/eli/dir/2017/541/oj/eng.
28 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the implementation of Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online* (Luxembourg: Publications Office of the EU, 2024), 10-11, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2024%3A64%3AFIN.
29 "The Digital Services Act: Ensuring a safe and accountable online environment," European Commission, accessed February 15, 2025, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.
30 On a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), regulation (EU) 2022/2065 European Parliament and Council of the EU art. 33 (2022), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

to request access to non-public information concerning content moderation on behalf of the platforms, and culminating in the establishment of the DSA Transparency Database, which collects information on the reasoning of content removal by platforms.[31] Early analyses suggest that the majority of moderation actions relate to ToS violations rather than violations of law, highlighting the wide discretion platforms maintain in content moderation.[32] Another novelty of the DSA is the mandatory systemic risk assessments for VLOPs and VLOSEs, which must evaluate the impact of their moderation practices, algorithms, and recommender systems on public discourse, civic engagement, and fundamental rights.[33]

These two instruments are supported by adjacent regulatory frameworks at the EU level. The Artificial Intelligence Act (AI Act, Regulation EU 2024/1689) reinforces the TCO and DSA by introducing risk-based requirements for AI systems used in content moderation, mandating systemic risk assessments for general-purpose AI tools.[34] The General Data Protection Regulation (GDPR, Regulation EU 2016/679) governs the lawful processing and retention of user data in moderation activities, including the conditions under which fully automated decision-making is permissible.[35] Finally, the Audiovisual Media Services Directive (AVMSD, Directive EU 2018/1808) supplements the regime by addressing video-sharing platforms and requiring proportionate measures against audiovisual content inciting violence or hatred.[36]

However, implementation challenges remain for both the DSA and the TCO. This relates in particular to ensuring proportionality in content moderation on behalf of the platforms, insufficient support for smaller platforms, and robust safeguards against violations of fundamental rights such as freedom of speech through content moderation. In chapter 5, the definitions will be further elaborated upon.

1.2.3 Role of Private Sector in Content Moderation

When it comes to the moderation of online content, researchers found that HSPs are best suited to take on this task, compared to other stakeholders. National but also regional authorities are often unable to identify specific users, and in case they do, they frequently face jurisdictional issues depending on the residency of the user to take meaningful judicial action.[37] Indeed, scholars agree that HSPs play a crucial role in governing public discourse online, merely because they hold a trove of users' data required for content moderation and are in charge of the spaces in which public discourse takes place online, most importantly, social media platforms.[38] Nevertheless, disagreement arises about how much leeway private, profit-oriented actors such as HSPs should be given in deciding on normative frameworks that ultimately shape public discourse and based on which infringements on the right to freedom of speech are being made.[39] Notably, these private actors have no legal obligation to ensure individuals' enjoyment of human rights. This obligation lies with the States. Private actors are only obliged to support a

31 Chap. III DSA.
32 Rishabh Kaushal et al., "Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database," *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024): 1124-1125, https://doi.org/10.1145/3630106.3658960.
33 Art. 34 DSA.
34 Laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), regulation (EU) 2024/1689 European Parliament and Council of the EU (2024), http://data.europa.eu/eli/reg/2024/1689/oj.
35 On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), regulation (EU) 2016/679 European Parliament and Council of the EU, https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.
36 Amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, directive (EU) 2018/1808 European Parliament and Council of the EU (2018), https://eur-lex.europa.eu/eli/dir/2018/1808/oj/eng.
37 Amy-Louise Watkin, "Developing a Responsive Regulatory Approach to Online Terrorist Content on Tech Platforms," *Studies in Conflict & Terrorism* (2023): 1, https://doi.org/10.1080/1057610X.2023.2222891.
38 Brenda Dvoskin, "The Illusion of Inclusion: The False Promise of the New Governance Project for Content Moderation," *Fordham Law Review* 39, no. 4 (2025): 1317 & 1332-1333, https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=6141&context=flr.
39 For a discussion of different approaches, see Dvoskin, "The Illusion of Inclusion," 1315-1375.

State in carrying out law enforcement actions.[40] Hence, the EU had initially taken the approach of relying on voluntary cooperation of HSPs in limiting the dissemination of terrorist content online.[41] This did not lead to the desired results, according to the EU and the Member States, as terrorist and other illegal content continued to be spread on the internet, and approaches between HSPs significantly varied. Thus, the EU decided to take a more regulatory approach, starting with the TCO, later DSA, and other tech regulations outlined above. This approach aims to provide common definitions and uniform approaches at least on a regional European level, as well as transparency over the actions taken by HSPs in content moderation and the assurance of minimum human rights safeguards.[42]

However, research found that in complying with regulations on the moderation of online content, HSPs face three obstacles, as they are (i) unaware and lacking content matter expertise to comply, (ii) do not have the human and technical resources and financial capacities available, or (iii) are merely unwilling to comply.[43] When talking about compliance, one must also note that compliance with regulatory frameworks is not merely achieved by developing policies on content moderation. These policies must also be operationalised. To ensure comprehensive compliance of HSPs with the relevant regulatory frameworks, such as the TCO and DSA, the above-mentioned obstacles thus need to be overcome. Among others, this can be achieved through public-private partnerships in building the relevant capacities for HSPs, through awareness raising and education on the applicable regulatory frameworks, and to a lesser extent, through sanction regimes to encourage compliance.

1.3 Problem Statement, Objectives and Scope of the Study

In the current situation, various definitions are used to classify online content, and the ToS of HSPs - as well as their methods of moderating online content - differ significantly. This creates an undesirable situation, where there is a risk that problematic and harmful content is not moderated when it should be in accordance with the legal obligations, or, at the other extreme, that content is removed in a way that infringes on freedom of expression.

Against this background, WODC, at the request of NCTV, commissioned the International Centre for Counter-Terrorism (ICCT) to conduct a study to examine whether a reliable assessment framework can be developed that enables online platforms to identify and moderate terrorist, extremist, and borderline content in a more systematic manner and in accordance with legal frameworks, while respecting the right to freedom of expression (mid-term impact). Such an assessment framework would aim to offer a set of clear indicators that function as a checklist to classify content, either before it goes online or after it has been posted. This assessment framework could – if deemed feasible and developed - allow HSPs to better detect which content is not allowed according to EU law whilst respecting the freedom of expression and other relevant human rights. Overall, the study aims to contribute to the development of more effective strategies to strengthen the resilience of online platforms and society against terrorist, extremist, and borderline content, thereby helping to counter online radicalisation (long-term impact).

Considering the ambiguity related to the scope of definitions and concepts, most notably regarding the concept of borderline content, the research team has deemed it necessary to deviate from the terminology used in the research request (Startnotitie), namely, terrorist, extremist and borderline content. Instead, **for the purpose of this study and to limit the scope,**

40 Rikke Frank Jørgensen, "When private actors govern human rights," in *Research Handbook on Human Rights and Digital Technology* (Cheltenham: Edward Elgar Publishing, 2019), 346-363.
41 *§§* 6 & 40 preamble TCO.
42 *§ 10 preamble* TCO.
43 Watkin, "Developing a Responsive Regulatory Approach to Online Terrorist Content on Tech Platforms."

it is considered more constructive to use the terms terrorist, illegal, and implicit extremist content that is harmful in the context of violent extremism and terrorism. Particularly related to the feasibility question of an assessment framework, it is extremely important to be able to consider whether certain content falls under the protection of freedom of expression. Restricting the freedom of expression is only possible if there is a legal basis, if it serves a legitimate aim and meets the substantive criteria of proportionality, necessity and adequacy. To be able to assess whether content falls under freedom of expression, it is crucial to rely on legal definitions and key principles in moving forward in assessing the feasibility of an assessment framework. Thus, instead of using the term extremist content as the second category, which is not defined in law, the research team opted to use the category of 'illegal content' as far as it is related to terrorism and violent extremism. Finally, considering the fact that the term 'borderline' is also not defined in law, and can mean something different to different stakeholders, the research team opted for the term 'implicit extremist content' to describe the category of content that is initially not recognised to fall under the first two categories due to the fact that the meaning is concealed also causing confusion regarding the legality of the content, yet the impact is considered harmful.

1.4 Working Definitions

For the purpose of this study, the following working definitions will be used:

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.[44] 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 is understood as:

'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 elaborate background and explanation of these definitions can be found in chapter 4.

44 Directive on Combating Terrorism.

1.5 Outline of the Study

In chapter 2 of this study, the research questions will be listed. Next, in chapter 3, the methodology used for this feasibility study is explained. In chapter 4, the legal frameworks and the scope of the definitions applicable to the qualification of the content will be elaborated. This is followed by chapter 5, in which the focus will be on the manifestation of terrorist, extremist, and implicit extremist content online, the developments regarding the way content is concealed, and the way technology is used to spread content. Next, the way detection and moderation of content is done by HSP is reviewed in chapter 6. In chapter 7, a pilot codebook is presented to test the feasibility of designing an assessment framework for online content. Finally, in chapter 8, conclusions will be presented answering the key research questions, and a set of recommendations will be offered.

2. Research Questions

In this study, the feasibility of the development of a reliable assessment framework that enables online platforms to identify and moderate terrorist, illegal, and implicit extremist content without infringing on the fundamental right to freedom of expression will be examined.

If it is not considered to be feasible to develop an assessment framework for such content, both the enabling factors and obstructive barriers will be identified. These barriers may be legal, economic, or practical in nature. Where possible, concrete suggestions will be provided on how these barriers can be addressed.

The study will address the following questions:

1. What features or combination of features in online content determine whether it constitutes terrorist, illegal or implicit extremist content on online platforms?
2. In what ways and to what extent can these features - either individually or in combination - be detected and identified in online platform content?
3. What can be said about the reliability of detection methods for terrorist, illegal or implicit extremist content on online platforms?
4. How does the reliability of detection methods for various types of harmful content on online platforms (terrorist, illegal or implicit extremist content) relate to the risk of incorrect moderation decisions by online platforms, which may result in violations of the fundamental right to freedom of expression?
5. With current knowledge and technological capabilities, is it possible to develop a valid and reliable interpretive framework for detecting and identifying both explicit terrorist and illegal content, as well as implicit extremist content on online platforms, without unjustly infringing upon the fundamental right to freedom of expression?
6. Under what conditions could the implementation of an interpretive framework for terrorist, illegal and implicit extremist content contribute to reducing both the dissemination of such harmful content and the potential radicalisation of internet users? If a valid and reliable interpretive framework is deemed unfeasible:
7. What are the legal, technical, and possibly other barriers that hinder the development of an interpretive framework for terrorist, illegal and implicit extremist online content? What conditions must be met to remove these barriers, and how feasible is this?

3. Methodology

For this study a mixed research methodology is used, combining desktop research, semi-structured interviews, a roundtable expert consultation, and qualitative content analysis. For the latter, we conducted manual scraping techniques, identifying relevant online content to illustrate our findings and to create a testing sample to assess the pilot codebook created to evaluate the feasibility of the design of an assessment framework for terrorist, extremist and borderline content online. A data protection impact assessment has been conducted regarding the online scraping, and where privacy issues could occur mitigating measures have been taken to prevent harm. Below we present a more elaborate outline of the different research methods used.

3.1 Desk Research, Interviews, and Expert Roundtable

3.1.1 Desk Research

A desktop literature review has been conducted to reflect the recent threat assessment reports and the state-of-the-art literature related to the increase of terrorist, extremist, and borderline content online. (Academic) literature, policy reports, legislation, and jurisprudence published before 1 June 2025 has been taken into account.

Desktop research was also done to review current legal and policy frameworks, relevant jurisprudence, and Terms of Service (ToS) of hosting service providers (HSPs). The outcomes of the desktop research, *inter alia,* informed the identification of key indicators for an assessment framework.

3.1.2 Semi-Structured Interviews

Question lists were prepared to guide the semi-structured interviews conducted. Considering different stakeholder groups were targeted, the question lists contained a generic set of questions and a set of questions tailored to the specific stakeholder groups. The question lists are annexed to this report (Annex 2). For the interviews, the following stakeholders were approached: representatives of Instagram, Reddit, Tiktok, the NCTV, ATKM, the Dutch police, the Dutch Public Prosecutor, the EU Internet Referral Unit (EU IRU), experts from NGOs working in the field, EU Directorate-General for Migration and Home Affairs (EU DG Home), and representatives of the German Federal Ministry of the Interior (BMI) and the Bundeskriminalamt (BKA) unit specialised in dealing with online content.

Eight stakeholders accepted our invitation. A list of the stakeholders or stakeholder groups interviewed is annexed to this report (Annex 1). Although, it was not possible to conduct interviews with all stakeholders approached, the research team managed to conduct interviews with representatives of most stakeholders' groups, except with representatives of the online platforms. Considering the fact that NCTV in its cooperation with HSPs is interested to discuss the manner in which they detect, and moderate harmful content, the complete lack of interest of the approached HSPs to cooperate for the purpose of this study with the research team by offering their insights, was extremely disappointing, if not concerning, as it does not reflect a responsible and accountable role from their side vis-à-vis this topic that has such a big impact on our societies. Ultimately an assessment framework – if feasible – aims to ensure more consistency in the detecting and moderation of terrorist, illegal and implicit extremist content whilst respecting the freedom of expression.

The interviews were recorded for transcription purposes only. References to the interviews in the report are anonymised.

3.1.3. Expert Roundtable

A roundtable expert meeting was conducted to discuss the challenges related to the development of an assessment framework. The focus was on questions dealing with the delimitation of what constitutes borderline content, and ways of detection and moderation of content online. The discussion focused specifically on the technical, legal and ethical considerations related to the feasibility and desirability of an assessment framework. The Chatham House Rule applied to the meeting. There were three main themes discussed, each with a set of questions that needed to be addressed. (See the Annex 3 for the agenda of the Roundtable expert meeting.) The three themes were:

1. Scope and definitions;
2. Detection and moderation methods; Can we identify obstacles versus smart solutions?;
3. Discussing the feasibility and desirability of an assessment framework to unequivocally detect borderline content.

The following experts and stakeholders participated in the hybrid discussion: representatives of the NCTV, a representative of the National Police, a representative of the Public Prosecutor, a representative of the ATKM, a representative of the EU IRU, representatives of the Scientific Oversight Board to this Study, Academics, and representatives of NGOs. For the roundtable expert meeting, representatives of online platforms were also invited, but they did not accept our invitation.

The outcomes of the discussions during the expert roundtable have informed the drafting of a working definition of borderline content (see chapter 4) and also provided insights regarding the answering of the key research questions. Were relevant anonymous references to the discussion during the roundtable expert meeting will be made in the report.

3.2 Qualitative Content Analysis

Mindful of the overall objective of this study, namely to consider the feasibility of setting up an assessment framework of terrorist, illegal and implicit extremist content, which is effective, operational, and respects the freedom of speech right, and as part of the methodology used, also open source intelligence (OSINT) research has been conducted. We therefore collected online posts from several accounts on three different online platforms in relation to two separate events during a set period of time. 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 (see below in paragraph 3.2.4).

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 terrorist, illegal or implicit extremist content on key platforms. The research question tailored to those aspects will be answered based on the outcomes of the desktop research.

The OSINT research served the following purposes:

1. To collect a sample of online posts to test the potential effectiveness, accuracy, and operationality of indicators identified for an assessment framework.
2. To collect a sample of online posts that can serve as examples for the reflection on the aim, modus, and techniques used by terrorists and extremists to spread online content (see chapter 6).

3. To assist in reflecting on the ways HSP conduct their detection and moderation approaches.

To conduct the OSINT research, different steps needed to be taken. The first step is related to the selection of relevant online platforms to conduct OSINT research. The second step entailed the selection of relevant events to delimit the scope of the scraping exercise, especially related to the time frame in which online content would be scraped. As a third step, relevant accounts needed to be identified on each of the selected platforms. Step four entailed the online scraping of the selected accounts during the selected time frame. For step five, the scraped content was saved in two separate databases: one containing all information of the content, including account name and other identifiable data; the second one containing anonymised data. During the sixth and final step, the anonymised scraped content was scored by the research team, making use of the pilot codebook (see chapter 7).

3.2.1 Selection of the Platforms

This study focused especially on what is in the public debate referred to as borderline content, an area that is largely under-researched. Hence, the OSINT research had to focus on platforms where such borderline content can be found. Furthermore, it was considered important to be able to reflect on both jihadist and right-wing extremist borderline content, and also with a certain relevance for the context in the Netherlands.

General considerations

Previous research found that when faced with content moderation on a platform, users first adapt by engaging with and creating more borderline content.[45] In the next step, they navigate to less moderated or (partially) encrypted platforms where they can engage with more explicit terrorist/extremist content and do not need to resort to borderline framings.[46]

It was also considered beneficial to look at platforms where borderline content could still be found online. YouTube and Facebook, for example, were found to identify and remove/lower the visibility of a lot of content through AI, including many false positives (over-blocking). It was therefore deemed less likely that borderline or explicit content would still be available on these sites. A look at previous research indicated that right-wing extremists (RWEs) in the Netherlands were particularly active on Facebook, YouTube, X (previously Twitter), and to a certain extent, also on Instagram.[47]

A more recent study on online radicalisation and propaganda also mentions Facebook, Twitter (now X), TikTok, Instagram, and YouTube as the most popular mainstream platforms for extremists, with Telegram and Discord being the more fringe but popular ones too.[48] Furthermore, we also kept in mind that platforms that belong to the same company, for example, Instagram and Facebook, both belonging to Meta, might show – to a certain extent – similarities in the policies and practices of content moderation.

45 Molas, "Alt-solutism."; Fielitz, and Ahmed, "It's not funny anymore."
46 Heather Williams et al., "The Online Extremist Ecosystem: Its Evolution and a Framework for Separating Extreme from Mainstream," *RAND Cooperation* (December 2021), https://www.rand.org/pubs/perspectives/PEA1458-1.html; Alexandra Siegel, "Online Hate Speech," in *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge: Cambridge University Press, 2020), 72, https://doi.org/10.1017/9781108890960; Aleksandra Urman, and Stefan Katz, "What they do in the shadows: examining the far-right networks on Telegram," *Information, Communication & Society* 25, no. 7 (2022): 904–923, https://doi.org/10.1080/1369118X.2020.1803946.
47 Timo Peeters et al., *Online extreemrechtse radicalisering: Handvatten voor een preventieve aanpak* (Utrecht: Verwey Jonker Instituut, 2022), https://www.verwey-jonker.nl/wp-content/uploads/2022/12/121650_Online-extreemrechtse-radicalisering.pdf.
48 Charlie Stoeldraaijers et al., *Radicale reclame op sociale media: Een onderzoek naar online rekrutering door en voor extremistische groepen* (Amsterdam: Nederlands Studiecentrum Criminaliteit en Rechtshandhaving, 2024), https://repository.wodc.nl/bitstream/handle/20.500.12832/3382/3417-radicale-reclame-op-sociale-media-volledige-tekst.pdf?sequence=1&isAllowed=y; On softening of tone, see Erin Saltman, and Micalie Hunt, "Borderline Content: Understanding the Gray Zone," *GIFTC* (2023): 7, https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf. On mainstream (moderated) platforms and less moderated/niche platforms used by extremists and their different behaviour on the different platforms, see Williams et al., "The Online Extremist Ecosystem."

Borderline content can be found across different categories of content, first and foremost in hate speech. Anticipating the chance that we will get limited interaction/response from the platforms, we also keep in mind that certain platforms have signed on voluntarily to the Code of Conduct on Countering Illegal Hate Speech+.[49] The implementation reports already provide interesting data that we can use. Companies that have signed on are Dailymotion, Facebook, Instagram, Jeuxvideo.com, LinkedIn, Snapchat, Rakuten, Viber, TikTok, Twitch, X, and YouTube. It is also helpful to take into account which companies are members of cross-industry collaboration platforms focused on terrorism and extremism online, such as the Global Internet Forum to Counter Terrorism (GIFCT) are included in their research: Instagram, TikTok, Discord, Snapchat, Facebook, and LinkedIn.

This study relies on the assumption that context is paramount in determining whether certain conduct is legally permissible or must be removed when qualified as terrorist or illegal content. Hence, automated data scraping was not a preferred option. Instead, all data is being scraped manually. This comes with certain ethical implications for researchers as well as platform users. The latter do not know that they are being studied, hence data and information relevant for the study can only be reproduced without revealing identifying information (see below for more information on the data protection impact assessment conducted for this study). The fact that these users are studied without their consent and awareness also restricts the extent to which the researchers can engage with them. Interaction is thus strictly prohibited.[50] We furthermore respect the do no harm principle in our research approach. This means that we do not only consider this principle when considering privacy/reputational issues for the target audience we are researching, but we also keep the safety of our staff in mind. This means that for OSINT research, we only do anonymous and contactless research. Conducting OSINT into content on platforms for which the researchers need to identify themselves and get vetted before access is allowed is therefore excluded. The various considerations taken into account are reflected in the assessment document of the platforms (Annex 7).

Scope

Due to the budget limitations of this project, the scope of the OSINT research is rather limited, and merely used as a sample to reflect on the feasibility of setting up an assessment framework. The scope is therefore limited to the assessment of content on only three platforms, and content posted within the timeframe of a limited number of weeks (see below).

The languages are furthermore limited to Dutch and English, and the geographical focus is the Netherlands. Due to budget limitations, the research can only focus on platforms on which OSINT is easily possible without deploying additional tools that come at a cost. Concerning the geographical scope of the Netherlands, one must also take into account that the relevance of a platform in the Netherlands is not only determined by the number of Netherlands-based users, but also by whether Netherlands-related terrorist/extremist/borderline content has been shared on these platforms in the past.

49 The Code of Conduct on Countering Illegal Hate Speech Online +, which was integrated into the DSA on January 20, 2025, aims to enhance the way online platforms address illegal hate speech. It builds on the 2016 Code of Conduct and strengthens compliance with EU and national laws. The Code provides specific guidance for Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) to mitigate risks, improve enforcement, and ensure transparency in their actions against illegal hate speech. The European Commission and the European Board for Digital Services assess its effectiveness, offering recommendations for better reporting, data sharing, and stakeholder cooperation. See "The Code of conduct on countering illegal hate speech online +," European Commission, published January 20, 2025, https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online.

50 See National Ethics Council for Social and Behavioural Sciences, *Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants* (The Netherlands: National Ethics Council for Social and Behavioural Sciences, 2018), https://nethics.nl/onewebmedia/CODE%20OF%20ETHICS%20FOR%20RESEARCH%20IN%20THE%20SOCIAL%20AND%20BEHAVIOURAL%20SCIENCES%20v2%20230518-2018.pdf; An update Code of Ethics has been released in April 2025, https://nethics.nl/onewebmedia/Nethics-Code-of-Ethics-digitaal.pdf.

Related to the budget limitations, the uniformity and consistency in research methodology, in other words, the methods of scraping, saving, anonymising, coding, and analysing, need to be as uniform as possible for reasons of efficiency. In addition, considering the fact that the objective is to get an understanding of the way platforms moderate the content, we compare their policies with our own observations, hence, an important aspect of our overall consideration in selecting the platforms also relates to the cooperative attitudes of the platforms.

Selection of platforms

Ultimately, three platforms have been selected for the contactless online research, namely Instagram, TikTok, and Reddit. Several criteria were used to come to this selection, including the kind of content available on the platforms (jihadist, RWE, and borderline content), the number of followers/users globally per month, the average age of users, the types of content available (text, images, videos, other), the relevance of the platform for the Netherlands, the legal registration location of the platform, the methods of detecting harmful content (manual, AI, or combination), who can take moderation decisions (platform only, flaggers, users, or combination), whether contactless research is possible, and whether it is easily accessible for OSINT research.

These criteria were considered in relation to a total of nine platforms. In addition to the three platforms ultimately selected, other platforms that were considered included: Snapchat, Facebook, Discord, YouTube, Telegram, and X (previously Twitter). In the annex an overview is provided on how all of the nine platforms score on the criteria mentioned before (Annex 5). Ultimately, the assessment of all criteria taken together led to the final selection.

The selected platforms were chosen to offer the researchers the opportunity to analyse jihadist (Instagram and TikTok), right-wing extremist (Instagram, Reddit, and TikTok) and borderline (Instagram, Reddit, and TikTok) content online. The combination of platforms chosen, furthermore, offered an opportunity to reflect on the moderation done by a variety of platforms that qualify as very large platforms globally, while being large, medium, or small in the European context, and highly to average relevant to the Netherlands. The combination thus allowed to reflect on the difference of detection and moderation in relation to how big the platform is.

Two platforms are legally based in Ireland (Instagram and TikTok), and one is based in the Netherlands (Reddit). All are thus subjected to EU legislation. In addition, for Reddit Dutch legislation also applies.

The platforms selected use different kinds of content, allowing for an assessment of the reliability of detection methods for different kinds of content. For Instagram, content is mainly images and text-based comments, with, in addition, a messenger function. For Reddit, text is mainly posted in a chat format, and a lot of memes are used. For TikTok, the content is mainly temporary video clips and text-based comments.

In relation to the average age of the users, we see that TikTok caters to a very young audience, whereas Reddit targets mid-twenties, and Instagram twenties to early thirties.

The means used to detect harmful content also vary per platform, offering some insights into the reliability of these methods. Instagram uses mainly AI; Reddit uses its own community, and a combination of automated and human moderators; and TikTok recently moved to a mainly AI-operated method to detect harmful content (80 percent). With regard to who takes the decision on the moderation of the content, Instagram and TikTok are in charge of those decisions, whereas Reddit relies both on the users and the platform itself. Especially for so-called subreddits, there are

moderators who are users of the subreddit who can moderate content in that specific subreddit. Depending on the size of the subreddit, that can be several people.

Furthermore, it is important that all platforms are easily accessible for OSINT research so they can be analysed within the timeframe of this study.

These platforms are also relevant in the Dutch context regarding harmful content. Previous research has shown that extremists of all ideologies target mainstream social media platforms such as Instagram or Reddit in their recruitment efforts, which also affects Dutch residents.[51] Reddit has legal representation in the Netherlands, making the relevant Dutch authorities responsible for handling takedown requests for the platform and ensuring that the Digital Services Act (DSA) is enforced. Moreover, all selected platforms currently have moderation policies that address harmful content, including borderline content.

Finally, all three platforms allow for contactless research. Considering the question of how accessible the platforms are for OSINT research, Reddit was considered most accessible. Instagram is considered medium in terms of accessibility. This is also the case for TikTok, however, OSINT research on TikTok is more complicated due to algorithm-driven content and limited search functionalities, which are often deleted or difficult to trace. To overcome this hurdle, the researchers have developed methods to extract user data and analyse profiles. For instance, utilising TikTok's application programming interface (API) and third-party tools, it was possible to enable the collection of user information and content for analysis.[52] Particularly, these last two criteria resulted in ruling out certain platforms, because contactless research was either not possible (Discord), or limited due to the fact that certain interesting groups were closed, and access would not be possible without making contact (Facebook). Additionally, OSINT research on these platforms was considered difficult (Facebook and X) or not possible at all (Telegram and Discord).

3.2.2 Identification of Triggering Events

To systematically identify accounts disseminating borderline content related to the Netherlands, a selection process was implemented based on significant triggering events. These events were chosen for their national relevance and potential to elicit problematic reactions within both right-wing extremist (RWX) and Islamist spheres, particularly those associated with hate speech. Based on these criteria, the following two events were selected:

1. **The Amsterdam riots (6-7 November 2024):** Violent clashes broke out following a UEFA Europa League match between *Maccabi Tel Aviv supporters and pro-Palestinian groups, leading to multiple injuries and arrests. The observation period was 6-20 November 2024.*

2. **The *White Lives Matter* projection (31 December 2022):** During the New Year's Eve celebrations, racist slogans were projected onto the Erasmus Bridge, sparking public outrage and condemnation. The trial started on 17 December 2024 and ended on 16 January 2025. The observation period for this incident was 1-14 January 2023 and 17 December 2024 – 30 January 2025.

These events served as inclusion criteria for identifying relevant accounts and provided the basis for the data collection.

51 Stoeldraaijers et al., *Radicale reclame op sociale media*.
52 "A Guide to OSINT Investigation and Research on TikTok," Steve Adams, Intelligence with Steve, published November 13, 2019, https://www.intelligencewithsteve.com/post/osint-tiktok.

3.2.3 Selection of Accounts

Identification of keywords

Accounts selected for the research were identified through a keyword-based search strategy. Keywords were initially identified through a preliminary media review of articles covering the two selected triggering events, conducted in both English and Dutch. This review allowed to establish an initial set of general terms associated with the two events under investigation.

To refine this list, an exploratory round of social media scraping was conducted in English and Dutch across various social media platforms. In addition to platforms selected for this research, Instagram, TikTok, and Reddit, this exploration included searches on other platforms such as Telegram, X (formerly Twitter), and 4chan in order to build the most consolidated list of keywords possible. This process facilitated the identification of additional relevant terms, including terminology commonly used within far-right and Islamist discourse.

The finalised list of keywords[53] was translated bidirectionally between English and Dutch, except in cases where no direct equivalent existed in the other language (e.g., *Palestinisering).* However, different spellings of the same terms – including intentional misspellings used by some users to evade detection (see '*Preliminary observations*' below) – were not included in the keywords. The endless number of variations made it impossible to include.

Identification of accounts

Once the keyword list was consolidated, a targeted search was conducted to identify three relevant accounts on each of the three selected platforms (nine accounts in total). These searches were carried out using a combination of keywords. For content related to the Amsterdam riots, the keyword combination included (1) a term associated with the Netherlands, (2) a term associated with Israel, and (3) a general term referring to the event. For content related to the *White Lives Matter* projection, a combination of (1) a keyword associated with Rotterdam or the Erasmus Bridge and (2) a general term referring to the event.

The eventual keyword-based search was conducted using both platforms' internal search function and Google's 'site:' search operator (e.g., instagram.com 'Keyword 1' 'Keyword 2'), set to the data collection period of the respective trigger event. Using Google's 'site:' operator helped overcome some of the shortcomings of the platforms' own search functions, such as the non-exhaustiveness on Reddit and Instagram, as well as the algorithmic bias on Instagram.

Inclusion criteria and final selection of accounts

The content of the accounts identified through the keyword search was then scraped. In line with the objectives of this OSINT research – i.e. to allow for piloting of indicators/markers that could assist in identifying extremist and borderline content – the research team specifically examined the presence of certain types of problematic content and narratives, which may qualify as borderline. To do this, it built on GIFCT's mapping of borderline content types,[54] and paid particular attention to identifying the presence of hate speech, glorification or incitement to violence, mis- and disinformation, conspiracy theories, violent graphic content, weaponry instructional material, or violent extremist symbols and slogans. Accounts displaying the highest volume of such content in relation to the selected events were prioritised for final data collection.

53 The list of keywords is on file with the researchers, and can be shared upon request.
54 Saltman, and Hunt, "Borderline Content," 6-7.

This resulted in the selection of nine accounts:

Table 1: Selection of accounts broken down by triggering events and platforms

	Instagram	TikTok	Reddit
Amsterdam riots	3	1	3
White Lives Matter	0	2	0

- Seven accounts identified through content related to Amsterdam riots, including:
  - Three on Instagram, one on TikTok, three on Reddit;
  - Three spreading RWE-inspired borderline content;
  - Two spreading Islamist-inspired borderline content;
  - Two other-ideological accounts spreading borderline content.

- Two accounts identified through content related to the *White Lives Matter* projection, including:
  - Zero on Instagram, two on TikTok, zero on Reddit;
  - Two accounts spreading RWE-inspired borderline content;
  - Zero account spreading Islamist-inspired borderline content.

3.2.4 Scraping of Content and Analysis

All content published by these accounts over a fixed period of time was systematically examined.

For the Amsterdam riots, this period covered the two weeks following the event (6 – 21 November 2024). For the *White Lives Matter* event, data was collected over two periods: included the two weeks following the Erasmus Bridge projection (1 – 14 January 2023), along with the period of the trial (17 December 2024 – 31 January 2025). The court trial period was included for the *White Lives Matter* event because this triggering event overall yielded more limited results, with less accounts being identified as sharing problematic content related to this event, and including this period allowed for the capture of a more recent content. This approach was not applied to the case of the Amsterdam riots due to the large number of individuals charged, and the scale of legal proceedings made it impractical to extend the data collection into this later period.

All posts issued during the set period of the selected accounts were collected, saved in a protected drive on a hard drive not connected to the cloud service, and only made available to the members of the research team. After the data collection phase, the data were anonymised. The data were stored in a separate protective hard drive, accessible only by the members of the research team. These data will be stored for a period of twenty years in accordance with the regulations set by WODC.

Based on the legal analysis and the literature review, the team identified key indicators for the identification of terrorist, illegal or implicit extremist content to be included in a so-called Pilot Codebook. Each post collected during the scraping phase was coded in the Pilot Codebook. For that, we used Excel. For each post, the following information was recorded: account name, handle, URL, ideological affiliation (far-right or Islamist), date, and a brief summary of the post. Subsequently, the post was scored on the indicators for the various categories of content. These data were only accessible to the researcher who conducted the scraping and the project coordinator.

3.3 GDPR Considerations and Ethical Framework

Data is collected from users of online platforms who post public content. These users may also include minors. The identifiable information that may be obtained through scraping includes the name, handle, and/or account name of the person posting the information. Information may also link to an organisation or mention membership in an organisation, or to a place of residence. It may also link to an email address, work, or other demographic information. Links to specific groups or pages and likes on posts can also yield identifiable information. Finally, it is also theoretically possible that the post itself could be classified as criminally suspect, and therefore, privacy-sensitive.

For these reasons, and in line with the General Data Protection Regulation of the EU, ICCT conducted a data protection impact assessment (DPIA)[55] and took action to anonymise data as much as possible, and limit access to the privacy-sensitive data. The selected references in this report to online posts to illustrate our arguments do not in any way reveal the identity of the individual who initially posted the content.

3.4 Limitations of the Study

Due to time and budget constraints, only three platforms are selected for further analysis with respect to their ToS, and for our own OSINT research. While careful considerations were made to select platforms that represent different target audiences (for instance the average age of users), differences in the kind of content available on the platforms (jihadist, RWE and borderline content), the number of followers/users globally per month, the, the types of content available (text, images, videos, other), and the relevance of the platform for the Netherlands, conclusions drawn on the basis of this analysis is nevertheless limited, and not per se reflecting the policies used within the sector of HSPs as a whole.

The incidents selected for the OSINT research were chosen for the likelihood that jihadist, RWE and borderline content would be posted by the users of the accounts selected. Reflecting on the usability of selected criteria for the detection of terrorist, illegal or implicit extremist content is therefore limited to these ideologies.

The accounts selected were mostly posting content in English or Dutch. However, on some occasions posts were done in Arabic, or reference or links were posted to Arabic content. However, none of the research team members speaks Arabic, and although translation tools can translate text messages, images of Arabic cannot be translated, and therefore not assessed (this was the case for two posts).

Also, due to time and budget constraints, we were only able to perform a light touch intercoder reliability exercise. The sample for intercoding was limited, and we only used a four eyes, instead of a six eyes principle. However, the purpose of our intercoding was merely limited to internal learning on the use of the indicators, and to get a sense of the reliability or interpretability of the indicators used. The coding, after all, did not serve to cast a final judgment on the qualification of the coded posts.

55 The DPIA is available with the research team, and can be requested for review.

4. Legal Framework and Scope of Definitions

Against the background of ever evolving terrorist and extremist activities in the online sphere, national governments and regional bodies such as the EU have taken steps to regulate certain online content. This section provides a chronological overview of key pieces of legislation applicable in EU Member States and assesses other legislative developments that inform the framework on combating terrorist and other illegal content online. The analysis in this chapter will feed into the analysis of the feasibility of an assessment framework, since the legal definitions inform the identification of indicators for this assessment framework. These definitions have furthermore been used to identify the indicators that the research team used in the pilot codebook, which was designed to test the feasibility of an assessment framework, making use of the scraped online content.

4.1 Legal Framework

4.1.1 Historical Benchmark: NetzDG

The German *Netzwerkdurchsetzungsgesetz* (Network Enforcement Act, NetzDG), which was applicable to social media companies with more than two million registered users in Germany, was adopted in 2017 and entered into full effect in January 2018.[56] It is considered to be the first law worldwide that provides obligations for social media companies on the moderation of online content and imposing fines in case of systematic failure.[57] Due to tight timeframes for the deletion of apparently illegal and potentially illegal content and fines connected to systematic failure in establishing processes to ensure the timely deletion of such content, critics feared that the relevant online platforms would delete more content than required when in doubt, leading to a so-called "overblocking."[58] Additionally, many feared that a strict moderation of comments on social media platforms would lead to self-censorship of users out of fear that their content could be deleted.[59] A 2024 study on the deletion of comments on Facebook and YouTube in France, Germany, and Sweden found that Germany not only had the highest rate of deleted content on both platforms, but also the highest rate of false positives, meaning content that was deleted, albeit being legally permissible.[60] The authors of the study concluded that this "overblocking" could be linked to the fines implied under the NetzDG. However, they were unable to make a clear causal link.[61]

Nonetheless, another study found that there were no significant signs of an increase in content deletion across different ideological backgrounds at Facebook, before, during the transition phase, and after the NetzDG entered into force.[62] This study could not find indicators for self-censorship of Facebook users after the entry into force of the NetzDG. On the contrary, the study concluded that users were more active and that the tone of comments got more negative than before the NetzDG was in force.[63] In addition to platform owners, users can also moderate the comments posted under their own posts. Thus, a direct impact on the deletion of Facebook comments and the NetzDG could not be definitely established.[64]

56 NetzDG, I no. 61 Bundesgezets.
57 Maaß et al., "Evaluating the regulation of social media," 4.
58 Maaß et al., "Evaluating the regulation of social media," 4-5.
59 Maaß et al., "Evaluating the regulation of social media," 5.
60 The Future of Free Speech, *Preventing "Torrents of Hate" or Stifling Free Expression Online? An Assessment of Social Media Content Removal in France, Germany, and Sweden* (Nashville: The Future of Free Speech, 2024), 57, https://futurefreespeech.org/wp-content/uploads/2024/05/Preventing-Torrents-of-Hate-or-Stifling-Free-Expression-Online-The-Future-of-Free-Speech.pdf.
61 The Future of Free Speech, *Preventing "Torrents of Hate" or Stifling Free Expression Online?* 50 & 53; Maaß et al., "Evaluating the regulation of social media," 5.
62 Maaß et al., "Evaluating the regulation of social media," 10.
63 Maaß et al., "Evaluating the regulation of social media," 10-12.
64 Maaß et al., "Evaluating the regulation of social media," 15.

Regardless of the only marginal effects of overblocking and self-censorship due to the NetzDG in Germany, the concerns about infringements on the freedom of speech by letting private actors moderate online discourses based on broad domestic laws remained[65] – albeit not only in the German context. A 2019 study has found that the law was used as a blueprint by several authoritarian states around the world in their efforts to censor the online space.[66] Notably, states acted on two critical pillars to amend the NetzDG as a tool for online censorship: they refer to broad categories of content and vaguely referred to criminal laws to identify content that must be removed, and they abolished judicial oversight of content moderation.[67] While the study concludes that democracies and regional organisations such as the EU should be more cautious in adopting content moderation laws to prevent abuse by less democratic states or parties,[68] it also pointed out that many of the countries already had strict online regulations in place that were violating freedom of speech and were only further tightened through inspiration from the NetzDG.[69]

Some important observations on how to improve future content moderation laws while protecting the freedom of speech can be drawn from these alleged abuses of the NetzDG. These key lessons for future legislative attempts on online regulation, in particular, relate to:

1. ensuring independent judicial oversight and easily accessible remedies for affected users,
2. clearly listing defined categories of removable content and allowing for less grave actions to ensure the protection of other rights violated by certain online content,
3. allowing for public scrutiny through transparency reporting.

The latter aspect is often overlooked when discussing the NetzDG, despite the law taking the novel approach of obliging social media platforms to publish biannual transparency reports on their moderation activities in case they have received more than 100 complaints about illegal content within one calendar year.[70] In fact, most studies assessing the impact of the NetzDG on freedom of speech online have relied on these reports,[71] thus allowing for independent review of moderation practices and creating public pressure to improve these practices.

4.1.2 EU Regulation on Terrorist Content Online

To limit the dissemination of terrorist content online, Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (TCO) provides several obligations for different stakeholders, including so-called hosting service providers (HSPs). Pursuant to the definition contained in Article 2(1) TCO, this includes any provider, regardless of whether it has an establishment in the EU or not, as long as there is "a substantial connection to that Member State or those Member States" in which they offer their services to "natural or legal persons in one or more Member States."[72] In this regard, the preamble of the TCO indicates that one must take into account all relevant factors, "including [...] the use of a language or a currency generally used in that Member State, or the possibility of ordering goods or services."[73] Since the TCO does not make any reference to the size of the

65 The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression to the Chancellor of Germany, June 1, 2017, https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf.
66 Mchangama, and Fiss, "The Digital Berlin Wall," 6; For a more recent, extensive list of countries adopting regulations based on the NetzDG, see Mchangama, and Alkiviadou, "The Digital Berlin Wall – Act Two."
67 Mchangama, and Fiss, "The Digital Berlin Wall," 17.
68 Mchangama, and Fiss, "The Digital Berlin Wall," 17.
69 Mchangama, and Fiss, "The Digital Berlin Wall," 8-9.
70 Para. 2 NetzDG.
71 Maaß et al., "Evaluating the regulation of social media," 2.
72 Pursuant to Art. 2 (5) TCO, such a substantial connection can arise either from "having a significant number of users [...] in one or more Member States" or from the fact that the provider is "targeting [...] its activities to one or more Member States". Other indicators for whether the service is targeted at one or more MS according to the preamble, could be "the availability of an application in the relevant national application store, [...] local advertising or advertising in the language used in that Member State, or [...] the handling of customer relations such as by providing customer service in the language generally used in that Member State."
73 § 16 preamble TCO.

HSP, it applies to all sizes of online companies who are providing an information society service, meaning "any service normally [but not necessarily] provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services," which consists of the storage of information provided by and at the request of a user, or public dissemination of information to a potentially unlimited number of people at the request of a user.[74]

The preamble of the TCO slightly expands the scope again by saying that when content is being shared among a group of users which is freely accessible without human discretion, content shall be considered to be disseminated to a potentially unlimited number of people.[75] Providers that offer a range of services, of which one is related to a group of users accessible only with human discretion, these providers still fall within the scope of the TCO as long as the other services provided relate to storage or public dissemination of user-provided content.[76] For all HSPs falling within the scope of the TCO, the regulation sets out six main obligations:

1. Removal or disabling of terrorist content within one hour after receiving a removal order by a competent authority (Article 3 TCO).
2. Taking specific measures with regard to content moderation if the service has been identified as being exposed to terrorist content (Article 5 TCO).
3. Preservation of removed or disabled content and related data for six months if the action resulted from a removal order or specific measures (Article 6 TCO).
4. Transparency about content moderation concerning terrorist content in the terms of service, and – in case of exposure to terrorist content or after having received a removal order within a calendar year – publication of annual transparency reports detailing measures to counter terrorist content (Article 7 TCO).
5. Provide user-friendly appeal mechanisms (Article 10 TCO).
6. Informing users about removed or disabled content. Reasoning and outlining remedies have to be provided only upon request (Article 11 TCO).

Consequently, HSPs that fall outside the scope of the TCO, for example messenger services with human discretion in admitting users to closed chats, may have less incentive to cooperate with competent authorities in expeditiously removing terrorist content or assist them in criminal prosecutions, to implement specific measures to reduce terrorist content on their platforms, and to be transparent about their content moderation, including available appeal mechanisms due to not being bound by the TCO.

Nonetheless, there are informal means to mitigate some of these effects. First and foremost, HSPs that are willing to work towards a transparent and human rights compliant content moderation can join networks with other tech companies or intersectoral fora to exchange good practices.[77] These HSPs can also use public materials, mostly issued by non-governmental organisations, providing guidance on content moderation.[78] Cooperation of HSPs with competent authorities in the removal of terrorist content can also take place through the EU Internet Referral Unit (EU IRU) at Europol, which can flag terrorist content on behalf of a Member State authority to any internet company providing services to EU citizens. These referrals are non-binding for the companies but allow them to assess the items against their own Terms of Service (ToS) and potentially delete them in case of a violation of these terms.[79] However, it is unclear how often such referrals actually lead to the deletion of content since the EU IRU's annual transparency reports do not

74 See Art. 2 (1) TCO in conjunction with Art. 1 (1b) AVMSD.
75 § 14 preamble TCO.
76 The President of the Council of the EU to the Council of the EU, 11.
77 The Global Internet Forum to Counter-Terrorism (GIFCT) for example provides a mentorship programme for providers who do not yet meet the network criteria such as having included terrorist content in their terms of service or issuing transparency efforts, see "Membership," GIFCT, accessed February 3, 2025, https://gifct.org/membership/.
78 Notably Tech Against Terrorism and the Global Internet Forum to Counter-Terrorism.
79 Europol, *2021 EU Internet Referral Unit Transparency Report* (Luxembourg: Publications Office of the EU, 2022), 7-9, https://www.europol.europa.eu/cms/sites/default/files/documents/EU_IRU_Transparency_Report_2021.pdf.

provide information on the removal rate of the several thousand referrals they send each year to more than a hundred internet companies.[80] Initial evidence suggests an increased incentive by HSPs to process referrals since the TCO entered into force to avoid receiving removal orders, which could give rise to additional obligations under the TCO as described above.[81] Experts interviewed for this study also confirmed that some Member States apply a scaling approach as they first rely on the non-binding referrals through the EU IRU and only escalate to mandatory removal orders under the TCO if the relevant HSP fails to adequately respond to the referral.[82] On the other hand, one must acknowledge that some HSPs, regardless of whether they fall within the scope of the TCO or not, are deliberately refusing to moderate terrorist and extremist content on their platforms and even promote themselves as opposing censorship and advocating for free speech.[83]

In defining what constitutes terrorist content in the first place, the TCO heavily relies on the definition of terrorist offences and group related terrorist offences in Articles 3 and 4 of the Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA which will be discussed below.[84]

One of the main criticisms of the TCO by experts is that it is posing unfair burdens on smaller HSPs who do not have the necessary resources and knowledge to implement all obligations at the speed that big tech companies can afford.[85] This relates in particular to the one hour deadline to process removal orders from competent authorities, which, in case of non-compliance can lead to harsh fines. While small HSPs hence prioritise building structures to expeditiously process removal orders under the TCO, as of December 2023, removal orders have only been issued to a dozen of HSPs, most of them big tech companies.[86] At the same time, removal orders are a trigger of additional obligations for HSPs under the TCO, most importantly the obligation to issue public transparency reports on ongoing content moderation efforts, available appeal mechanisms, and data on removal orders pursuant to Article 7 TCO.

To support small and medium-sized HSPs in their implementation of the TCO, the European Commission has assigned three separate projects (FRISCO, ALLIES, and TATE) that not only seek to raise awareness of the TCO and provide a platform to share good practices in content moderation, but also to encourage small HSPs to deploy automated content detection and moderation tools.[87]

4.1.3 Digital Services Act

Only one and a half year after the TCO had been adopted, another piece of EU legislation seeking to address the issue of online content, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Framework for the Digital Services Act (DSA), was adopted. Since the scope of content the DSA is addressing is wider than only terrorist content, many see the TCO as *lex specialis* to the DSA. The latter was adopted with the goal to harmonise the digital market across all Member States with a view to consumer protection, reduction of harmful content, and increased accountability of service providers.[88]

80 For the latest numbers of 2023, see Europol, *2023 EU Internet Referral Unit Transparency Report* (Luxembourg: Publications Office of the EU, 2025), 6, https://www.europol.europa.eu/cms/sites/default/files/documents/2023%20EU%20Internet%20Referral%20Unit%20Transparency%20Report.pdf.
81 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL*, 2, 10 & 13.
82 Interview conducted on 27 June 2025, on file with research team.
83 Williams et al., "The Online Extremist Ecosystem," 3-4, 8 & 24.
84 Directive on Combating Terrorism.
85 Giovanni Buttarelli, *Formal comments of the EDPS on the Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online* (Brussels: EDPS, 2019), https://www.edps.europa.eu/sites/default/files/publication/2018-02-13_edps_formal_comments_online_terrorism_regulation_en.pdf.
86 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL*, 6.
87 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL*, 10-11.
88 Aina Turillazzi et al., "The Digital Services Act: An Analysis of Its Ethical, Legal, and Social Implications," *Law, Innovation and Technology* 15, no. 1 (2023): 86 & 88, https://www.tandfonline.com/doi/full/10.1080/17579961.2023.2184136?src=recsys.

To this end, the DSA poses obligations on a variety of stakeholders providing their services within the EU single market, regardless of whether they have an establishment within the EU or not. It does so by providing sub-categories of stakeholders which have different obligations. All of the relevant stakeholders are considered "intermediary services" pursuant to Article 3 lit. g DSA. These are then broken down into three further sub-categories: (1) hosting service providers, (2) online platforms, which are considered a sub-category of hosting services, and (3) very large online platforms and search engines, which are a sub-category of online platforms.[89] According to this pyramidal categorisation, those HSPs with more than 45 million monthly users in the EU, which qualify as very large online platforms (VLOPs) and very large online search engines (VLOSEs)[90] bear the most responsibilities under the DSA in three key areas: transparency (including data access for researchers), oversight mechanisms, and efforts to counter illegal and harmful goods, services, and content.[91] Concerning hosting service providers' efforts to limit certain online content, the DSA moves one step further than the TCO in two regards.

Firstly, the DSA casts a wider net in terms of unwanted content than the TCO. While the TCO referred to a specific set of terrorist offences in line with existing EU legislation, the DSA refers to illegal content. While that in many cases includes terrorist content, it can go beyond that and relate to other offences such as incitement to hatred, or defamation, as it can contain any violation of national or Union law, "irrespective of the precise subject matter or nature of that law."[92] This wide scope of illegal content forms the basis of the vast obligations on hosting service providers in relation to their content moderation efforts under the DSA. These include, for example, amendments to ToS and code of conduct where necessary, and action on content that was flagged by competent authorities or trusted flaggers.[93] The latter obligation is of particular concern to HSPs, as this constitutes an exemption of the general rule that HSPs cannot be held liable for the content provided by others. However, by obliging them to act upon flagged content, they can be held liable for not taking action and face stricter regulations on how to act.[94] This could ultimately result in over-blocking or -deletion of potentially illegal content by HSPs as they try to already remove as much content as possible that could be later flagged to eventually evade the obligation to duly act upon request.[95] Research has further shown that several platforms design their user reporting mechanisms in a way to avoid administrative burden and human intervention that comes with following the reporting procedures under Article 16 DSA, but also avoiding costly dispute settlement under Articles 20 and 21 DSA that can result from not acting upon content reports based on the DSA.[96]

Secondly, the DSA requires VLOPs and VLOSEs to conduct risk assessments. Through these assessments, VLOPs and VLOSEs, among others, shall examine how their content moderation systems, algorithmic and recommender systems, and the enforcement of their ToS relate to so-called "systemic risks". These risks are not only related to the spread of illegal content or goods but also "negative effects on civic discourse and electoral processes, and public security," or wider negative effects on individuals' mental and physical well-being.[97] Risk assessments constitute a novel approach in regulating the online sphere and are thought to complement existing fundamental rights safeguards, which merely focus on individual users and are often insufficiently implemented.[98] One systemic risk could be how users are evading content moderation by resorting to borderline content. Nonetheless, this has not received much attention

89 "The Digital Services Act," European Commission.
90 Art. 33 DSA.
91 For a breakdown of individual obligations under the key areas, see Turillazzi et al., "The Digital Services Act," 87.
92 Art. 3 (h) DSA.
93 Art. 22 DSA.
94 Art. 20-21 DSA.
95 Turillazzi et al., "The Digital Services Act," 93-94.
96 "Follow Me to Unregulated Waters! Are Major Online Platforms Violating the DSA's Rules on Notice and Action?" Daniel Holznagel, Verfassungsblog, published May 30, 2024, https://verfassungsblog.de/follow-me-to-unregulated-waters/.
97 Art. 34 DSA.
98 Niklas Eder, "Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation," *German Law Journal* (2024): 1198, https://doi.org/10.1017/glj.2024.24.

in the first round of risk assessment reports issued by VLOPs and VLOSEs in 2024. Instagram, for example, has merely mentioned borderline content once in their assessment with regards to hate speech, noting that it is used to evade content moderation but stating that the platform currently has no policies or practices to tackle the phenomenon systematically.[99] TikTok did not mention borderline content in its 2023 risk assessment at all.[100] The little attention given to borderline content as an emerging tactic to circumvent content moderation in the platforms' risk assessments might also be due to the fact that these assessments do not all cover the most recent reporting period.[101] Additionally, it has been criticised that these risk assessments are audited by consultancy and law firms, while they only provide little access to relevant data for civil society organisations that attempt to examine the assessments, for example, in relation to the risk of borderline content and suitable mitigation measures.[102]

Another key obligation for hosting service providers under the DSA certainly stems from the TCO, namely, transparency. Under the DSA, hosting service providers not only have to be transparent about their ToS and advertising mechanisms, but also their content moderation efforts in general and the reasoning in individual cases of content take down.[103] The latter presents a novel approach to ensuring platform transparency. VLOPs and VLOSEs (and with a transition period of one year, all other hosting service providers engaged in content moderation) are required to submit anonymised statements of reasons (SoR) for action taken on illegal content or content that violates the service's ToS to the European Commission without undue delay.[104] The Commission, in turn, is responsible to manage a "publicly accessible machine-readable database" in which these anonymised statements of reason are compiled.[105] To this end, the Commission has established the openly accessible DSA Transparency Database, which tracks the content moderation decisions of HSPs and allows users to analyse the mechanisms and reasoning used to take these decisions.[106] However, a first comprehensive analysis of the database as of November 2023 shows that these statements of reasons remain insufficient to provide proper transparency about content moderation decisions taken by hosting service providers. Researchers found that the template provided by the European Commission leaves too much uncertainty and leverage to HSPs regarding the type of information and level of detail they are providing.[107]

4.1.4 Artificial Intelligence Act

As outlined above, under the TCO and the DSA HSPs are allowed and arguably encouraged to deploy artificial intelligence (AI) tools to moderate terrorist and other illegal content.[108] Additionally, law enforcement agencies across Europe and at the regional level are also using AI tools to quickly identify terrorist and other illegal content online.[109] In this vein, the Permanent Assembly of the Organization for Cooperation and Security in Europe (OSCE) as part of the Bucharest Declaration and Resolutions encouraged States to invest in the development of AI tools for "detecting, monitoring, and countering terrorist activities, while ensuring transparency, accountability, and adherence to human rights and ethical standards in the design and implementation of such

99 Meta, *Regulation (EU) 2022/2065 Digital Services Act (DSA): Systemic Risk Assessment and Mitigation Report for Instagram* (Menlo Park: Meta, 2024), 74-75, https://transparency.meta.com/sr/dsa-sra_results_report-2024-instagram.
100 TikTok, *DSA Risk Assessment Report 2023* (Dublin: TikTok, 2023), https://panoptykon.org/sites/default/files/2025-01/tiktok-dsa-risk-assessment-report-2023.pdf.
101 "DSA risk assessment reports: A guide to the first rollout and what's next," John Albert, DSA Observatory, published December 9, 2024, https://dsa-observatory.eu/2024/12/09/dsa-risk-assessment-reports-are-in-a-guide-to-the-first-rollout-and-whats-next/.
102 Eder, "Making Systemic Risk Assessments Work."
103 Chapter III DSA.
104 Art. 17 DSA.
105 Art. 24 (5) DSA.
106 "DSA Transparency Database," European Commission, accessed September 10, 2025, https://transparency.dsa.ec.europa.eu/.
107 Kaushal et al., "Automated Transparency," 1123.
108 This is also acknowledged by later EU legislation on AI tools, see AI Act, regulation (EU) 2024/1689 European Parliament and Council of the EU § 120 preamble, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
109 For an overview of various AI technologies used by law enforcement, see e.g. Europol Innovation Lab, *AI and policing: The benefits and challenges of artificial intelligence for law enforcement* (Luxembourg: Publications Office of the EU, 2024), https://www.europol.europa.eu/cms/sites/default/files/documents/AI-and-policing.pdf.

technologies, including its use by law enforcement."[110] While this resolution is non-binding on the OSCE participating States, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on AI provides binding provisions for EU Member States to implement in their domestic laws. It takes a risk-based approach to the development and deployment of AI tools by the private sector and public authorities. In categorising AI tools in different risk categories, depending on their use, the AI Act sets out different requirements and obligations for developers and deployers, depending on the risk category of the AI technology.[111]

This also has an impact on AI-based content detection and moderation by HSPs. Similar to the DSA, the AI Act mandates users of certain AI mechanisms to conduct assessments of systemic risks that arise from this use.[112] This also includes HSPs that use AI tools in content detection, and moderation since they deploy "general purpose AI models" as defined in Article 3 (63) AI Act which have "high impact capabilities" as they perform among the most advanced general purpose models due to the large amount of data they are operating with and the possibility to integrate them into other systems.[113] In fact, experts argue that most online related AI systems on the market currently have high impact capability and hence require risk assessments by their developers.[114] If developed by the HSPs in-house, these assessment obligations hence fall on the HSPs themselves. If the tools are developed by external companies, these companies must conduct the risk assessment.

While the definition of systemic risks in Article 3 (65) AI Act slightly differs from the one used in Article 34 DSA, it is equally vague by stating that any risks that are "specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" are considered systemic. To reduce the burden of VLOPs and VLOSEs who are obliged to conduct risk assessments under the DSA, paragraph 118 of the AI Act's preamble provides that these risk assessments are considered to fulfil the risk assessment duties under the AI Act as long as they also examine AI tools used by HSPs.

However, there are still competing competencies concerning the enforcement of the risk assessment obligations under the DSA and the AI Act. For the DSA, VLOPs and VLOSEs have to report their assessments to the Digital Service Coordinator in their Member State of establishment or registration.[115] Under the AI Act, however, risk assessments must be reported to the EU Commission's Artificial Intelligence Office.[116] Thus far, it remains unclear whether VLOPs and VLOSEs have to communicate to the EU AI Office separately, whether their reports or relevant parts thereof will be shared with the Office by the national Digital Service Coordinators, or if there is any communication with the AI Office on the risk assessments of VLOPs and VLOSEs at all. It must further be noted that the risk assessment obligations under the DSA only apply to VLOPs and VLOSEs, smaller HSPs could thus fall under the AI Act if they develop recommendation, detection, and moderation tools in-house. This would create similar concurring reporting competencies in cases where these smaller HSPs are designated by the competent national authority as a HSP that is exposed to terrorist content under Article 5 TCO, pursuant to which they also have to report to the competent authority on their use of "specific measures" to reduce

110 Resolution on Artificial Intelligence and the Fight Against Terrorism, 31 OSCE Permanent Assembly § 21, https://www.oscepa.org/en/documents/ad-hoc-committees-and-working-groups/ad-hoc-committee-on-countering-terrorism/resolutions-and-publications/5040-resolution-on-artificial-intelligence-and-the-fight-against-terrorism-adopted-at-the-31st-annual-session-bucharest-29-june-to-3-july-2024/file.
111 AI Act.
112 Art. 56 AI Act.
113Art. 3 (64) AI Act.
114 Deborah Yao, "EU AI Act Would Scrutinize Many 'General' AI Models – SXSW 2024," *AI Business*, March 13, 2024, https://aibusiness.com/responsible-ai/eu-ai-act-would-scrutinize-many-general-ai-models-sxsw-2024#close-modal.
115 Art. 34 DSA.
116 "European AI Office," European Commission, accessed March 4, 2025, https://digital-strategy.ec.europa.eu/en/policies/ai-office.

the dissemination of terrorist content on their platforms. However, as of 31 December 2023, only one HSP in Germany was classified as being exposed to terrorist content under Article 5 TCO. The European Commission reported that for the same date, there were no other HSP with this classification in another Member State.[117] Hence, the additional reporting for smaller HSPs on their use of AI detection and moderation tools does not seem to be of much practical use as of early 2025.

4.1.5 General Data Protection Regulation

When moderating online content, HSPs must use, process, and collect data from users. Under certain conditions, they must even store data for several months to facilitate criminal investigations.[118] Thus, another piece of EU legislation is relevant to content moderation online, namely Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), the so-called GDPR.

Notably, Article 22 (2) GDPR generally prohibits the use of personal data for fully automated decision-making if that leads to "legal effects concerning [the data subject]". Such effects are arguably taking place when moderating online content, as it can interfere with users' right to freedom of expression as outlined below. Nonetheless, Article 22 (2) lit. b GDPR provides an exemption that is also relevant to the moderation of online content as it provides that automated decision making is permitted as long as it is in compliance with Union of Member State's law and provides sufficient safeguards to protect the rights of the data subject. Since both TCO and DSA provide HSPs with the option to use automated content detection and moderation as long as they are transparent about it, such automated decisions may be allowed under the GDPR. However, unlike the two other exceptions to automated decision-making laid out in Article 22 (2) GDPR, no human intervention is required if the automated decision-making, for example, content moderation, is in line with existing Union and Member States' laws and provides for other safeguards. Hence, according to the GDPR, HSPs are allowed to conduct fully automated content moderation without human intervention as long as they adhere to the safeguards of the TCO and DSA, which in turn mostly relate to transparency and accessible remedies as explained above.

The GDPR itself also includes some provisions on transparency. Notably, Articles 12, 23, and 14 GDPR lay out the obligations of data controllers, such as HSPs, on when and how to inform data subjects, meaning users of their services, about the processing of their data. When personal data is processed in content moderation, for example, through automated detection and moderation tools, this must be disclosed to the users.

Finally, a cornerstone of the GDPR is the provision on the so-called "right to be forgotten". Pursuant to Article 17 (1) GDPR, users are allowed to request the deletion of their personal data, which can also include content they have posted. This deletion does not only relate to the frontend of an online service, meaning that the data is no longer visible to the outside world. It also relates to the data processors, for example, an HSP's, internal data storage. However, Article 17 (3) already provides certain exceptions in which the storage of personal data is allowed, even without the consent of the data subject. Notably, this is the case when the data is still needed for the exercise of other EU or Member State's legal obligations. For example, Article 6 (1) TCO obliges HSPs to store personal data and other data related to content that has been deleted or made inaccessible for up to six months to support potential appeals as well as the "prevention, detection, investigation and prosecution of terrorist offences." This is not only related to content

117 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL*, 8.
118 Art. 6 Directive on Combating Terrorism.

that has been removed or disabled following a removal order, but also through the use of specific measures, for example (partially) automated or human content moderation, carried out in compliance with Article 5 of the TCO. However, both the TCO and the GDPR remain silent on the modalities of data storage in cases where the content moderation practices that led to the deletion of the content were not mandated by the TCO. The DSA, on the other hand, refers to the GDPR in its preamble, saying that all data processed for the purposes of compliance with the DSA must be handled in accordance with the GDPR, including the provisions on storage and deletion of personal data.[119]

4.1.6 Audiovisual Media Services Directive

Lastly, the Audiovisual Media Services Directive (AVMSD) plays a role in the moderation of online content, in particular where such content is not solely text-based but comes in the form of videos in view of changing market realities.[120] As the AVMSD is targeted at video-sharing platforms, it has an impact on several HSPs that, as part of their services, also offer users to share video content on which the HSP does not have any editorial influence.[121] Amending existing EU legislation on audiovisual media services, the 2018 Directive sets additional obligations for video sharing platform services. In relation to potential moderation of content, Article 28b AVMSD provides that all video-sharing platform services shall take "appropriate measures to protect [...] the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on [discriminatory grounds]."

While the AVMSD provides for Member States' control of appropriate measures, it leaves the video-sharing platform services considerable leeway in determining what exact measures they apply, by outlining that "measures shall be practicable and proportionate, taking into account the size of the video-sharing platform service and the nature of the service that is provided."[122] Unlike the TCO and DSA that encourage content moderation and proactive content detection and decision making by the platforms themselves, the AVMSD takes a softer approach in Article 28b (3), relying on users to flag and report content, refine their terms and conditions to allow for the removal of content, provide for appeals mechanisms, and foster media literacy.

This approach to limiting the dissemination of illegal, including terrorist content, bears less risk of infringing on the fundamental rights of content creators. However, the AVMSD puts emphasis on the role of Member States in observing implementation of the Directive by video-sharing platform services, which leads to a more fragmented approach than the TCO and DSA that provide for more coordination among competent Member States authorities.[123] Nevertheless, the AVMDS presents an important addition to the more narrow obligations of the TCO and DSA for HSPs that provide video-sharing platform services and are established in one or more Member States when attempting to limit the dissemination of terrorist, extremist, and other illegal content on their platforms.

4.2 Scope of Definitions

While the previous section provided an overview of the regulatory framework and the obligations imposed on different online service providers, this section will take a closer look at the different types of online content and how they are defined by different stakeholders. From a rule of law

119 *§§* 10 & 34 preamble DSA.
120 AVMSD.
121 Art. 1 (1b) AVMSD.
122 Art. 28b (3) AVMSD.
123 For illustration of differing approaches between Member States in applying the AVMSD to video-sharing platform services, see e.g. European Commission, *Reporting on the application of Directive 2010/13/EU "Audiovisual Media Services Directive" as amended by Directive (EU) 2018/1808, for the period 2019-2022* (Brussels: European Commission, 2024), 13-16, https://digital-strategy.ec.europa.eu/en/library/commission-report-application-audiovisual-media-services-directive.

perspective, definitions need to be clear in order to create certainty and foreseeability, both for the users as well as for the service providers. Hence, this section will critically examine the scope of definitions of illegal content, terrorist content, extremist content, borderline content, and harmful content.

4.2.1 Illegal Content

The DSA uses a very broad definition of illegal content as being "any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with EU law or the law of any Member State which is in compliance with EU law, irrespective of the precise subject matter or nature of that law".[124] In the recital of the DSA, further reference can be found to the concept of "'illegal content' [which] should be defined broadly to cover information relating to illegal content, products, services and activities. In particular, that concept should be understood to refer to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech or terrorist content and unlawful discriminatory content, or that the applicable rules render illegal in view of the fact that it relates to illegal activities."[125]

Illegal content *strictu sensu* refers to online content that is illegal by national or European law. Content may be illegal because it is illegal by itself or because it infringes on the consumer protection laws, or constitutes a violation of intellectual property rights. The guiding principle is that what is illegal offline is also illegal online. These definitions are relevant in legal proceedings and have been adopted by the platforms to moderate online content.

Under Dutch law, there is another distinction which relates to unlawful and illegal. While 'unlawful' refers to violations of rights under civil law,[126] 'illegal' refers solely to acts that constitute a criminal offence as defined in criminal law. Hence, not all unlawful acts are illegal and thus constitute a criminal offence, but all criminal offences are *ipso facto* considered unlawful. Spreading disinformation can be unlawful under Dutch law, but does not necessarily constitute a criminal offence. Similarly, sharing child sexual abuse material or inciting terrorism online are illegal and constitute criminal offences, but are also unlawful because they constitute a breach of a statutory duty. However, when the term 'harmful but lawful' or 'awful but lawful' is being used, it does not necessarily refer to the legal distinction between unlawful and illegal.

4.2.2 Terrorist Content

Under Regulation (EU) 2021/784, terrorist content includes any material that:

- **Incites or solicits** someone to commit or contribute to terrorist offences.
- **Solicits participation** in activities of a terrorist group.
- **Glorifies terrorist activities**, including by sharing material depicting terrorist attacks.
- **Provides instructions** on making or using explosives, firearms, or other weapons, including chemical, biological, radiological, or nuclear substances. [127]

When making a reference to a 'terrorist offence', the TCO Regulation aligns with existing EU law by referring to the set of terrorist offences in Article 3 of the EU Directive 2017/541:

(a) attacks upon a person's life which may cause death;

124 Art. 3 (h) DSA.
125 § 12 preamble DSA.
126 Algemeen gedeelte van het verbintenissenrecht, 6 Burgerlijk Wetboek art. 162 (1992), https://wetten.overheid.nl/BWBR0005289/2017-09-01/0/Boek6/Titeldeel3/Afdeling1/Artikel162/afdrukken.
127 Directive on Combating Terrorism.

(b) attacks upon the physical integrity of a person;
(c) kidnapping or hostage-taking;
(d) causing extensive destruction to a government or public facility, a transport system, an infrastructure facility, including an information system, a fixed platform located on the continental shelf, a public place or private property likely to endanger human life or result in major economic loss;
(e) seizure of aircraft, ships or other means of public or goods transport;
(f) manufacture, possession, acquisition, transport, supply or use of explosives or weapons, including chemical, biological, radiological or nuclear weapons, as well as research into, and development of, chemical, biological, radiological or nuclear weapons;
(g) release of dangerous substances, or causing fires, floods or explosions, the effect of which is to endanger human life;
(h) interfering with or disrupting the supply of water, power or any other fundamental natural resource, the effect of which is to endanger human life;
(i) illegal system interference and, if applicable, forms of illegal data interference under EU Directive 2013/40;
(j) threatening to commit any of the acts listed in points (a) to (i).

As outlined throughout this study, the incitement to commit terrorist offences, in particular glorification of terrorist acts in the online sphere, can pose significant challenges for both platforms and users, as the broad scope of glorification to terrorism can lead to unlawful restrictions of the freedom of expression. The reference to the glorification of terrorism in Article 1 (7) TC, modelled after Article 5 of the EU Counter-Terrorism Directive, raises several concerns.[128] Notably, it is being criticised that this offence neither requires a specific intent nor establishes a sufficient causal link between the act and the actual likelihood of a terrorist act.[129] Nevertheless, a draft bill that inter alia criminalises glorification of terrorism is under consideration in the Netherlands.[130]

4.2.3 Extremist Content

Similar to the term terrorism, there is no universally accepted definition for violent extremism or extremism. In the Netherlands, the Dutch Intelligence Service (AIVD) distinguishes between terrorism, extremism, and activism. With regard to extremism, the AIVD distinguishes between two forms:

1. The use of undemocratic methods that can undermine the functioning of the democratic legal order, and
2. Actively pursuing and/or supporting a change that in itself poses a threat to the democratic legal order, possibly through the use of undemocratic methods.[131]

The underlying activities can be violent or non-violent, and even though extremism as such is not criminalised in the Netherlands, several of the extremist acts can constitute a criminal offence,

128 Art. 5 Directive on Combating Terrorism. Public provocation to commit a terrorist offence: Member States shall take the necessary measures to ensure that the distribution, or otherwise making available by any means, whether online or offline, of a message to the public, with the intent to incite the commission of one of the offences listed in points (a) to (i) of Article 3(1), where such conduct, directly or indirectly, such as by the glorification of terrorist acts, advocates the commission of terrorist offences, thereby causing a danger that one or more such offences may be committed, is punishable as a criminal offence when committed intentionally.
129 Ilya Sobol, "Glorification of Terrorist Violence at the European Court of Human Rights," *Human Rights Law Review* 24, no. 3 (2024), https://academic.oup.com/hrlr/article/24/3/ngae017/7696469.
130 "Ministerraad stemt in met wetsvoorstel om verheerlijken van terrorisme strafbaar te stellen," Rijksoverheid, accessed September 10, 2025, https://www.rijksoverheid.nl/actueel/nieuws/2025/06/20/ministerraad-stemt-in-met-wetsvoorstel-om-verheerlijken-van-terrorisme-strafbaar-te-stellen. There is a lot of criticism on this draft proposal. See for instance the assessment by the Commission Meijers: https://www.commissie-meijers.nl/nl/comment/cm-310725-commissie-meijers-reactie-op-het-wetsvoorstel-strafbaarstelling-verheerlijken-van-terrorisme-en-openbare-steunbetuiging-aan-terroristische-organisaties/
131 "Wat is het verschil tussen activisme, extremisme en terrorisme?" AIVD, accessed September 10, 2025, https://www.aivd.nl/onderwerpen/extremisme/vraag-en-antwoord/wat-is-het-verschil-tussen-activisme-extremisme-en-terrorisme.

such as vandalism, incitement to hatred, or use of violence against persons.
In some other countries, such as the United Kingdom (UK), the government has adopted a definition of extremism. According to the UK definition, extremism is the promotion or advancement of an ideology based on violence, hatred or intolerance that aims to:

1) negate or destroy the fundamental rights and freedoms of others; or
2) undermine, overturn or replace the UK's system of liberal parliamentary democracy and democratic rights; or
3) intentionally create a permissive environment for others to achieve the results in (1) or (2).[132]

The undermining of the democratic legal order can be done through violent and non-violent activities. Examples of non-violent activities include systematically inciting hatred, spreading fear, disseminating disinformation, demonising and intimidating others, rejecting laws and regulations, and attempting to establish a parallel society in which the authority of the Dutch government and legal system is rejected.

A growing concern in the Netherlands is anti-institutional extremism. According to the AIVD, especially the narrative stating we are ruled by a 'malicious elite' is considered extremist as it undermines the democratic legal order. However, legally speaking, not all extremist or radical content constitutes terrorist content or other illegal content and it could be protected by the freedom of expression, no matter how harmful. The AIVD recognises that, in particular, in the context of anti-institutional extremism, the extremist narrative is sometimes being combined with lawful criticism.[133]

While all acts of terrorism constitute a criminal offence, not all acts of extremism are criminal offences.

As shown below, both terrorists and extremists use the internet and social media to - amongst others - spread their ideology. However, extremist content – unlike terrorist content – is not regulated by the TCO. The Dutch government, in December 2023, adopted the Enhanced Approach Online for Extremist and Terrorist Content, which describes extremist content and the concerns related to spreading inflammatory content that normalises extremist ideology whilst acknowledging the need to respect freedom of speech.[134] The Enhanced Approach underlines the difficulties in defining what constitutes extremist content, but at the same time acknowledges the need to develop guidance.

Example
The 764 is an online network that operates at the intersection of violent extremism, child sexual abuse material (CSAM) and other extreme violence such as animal cruelty, murder, and self-harm. Children are recruited through grooming and sextortion in online fora, gaming platforms and messaging apps. Recognising and understanding the linkages between terrorist content, illegal content, and harmful content that is being disseminated by 764 can be challenging. Moreover, different regulations apply to these different types of content, and online platforms have different policies for these categories of online content. Hence, an effective approach to counter such intersectional movements is widely missing at the moment.

132 "New definition of extremism (2024)," UK Government, published March 14, 2024, https://www.gov.uk/government/publications/new-definition-of-extremism-2024/new-definition-of-extremism-2024.
133 AIVD, *Anti-institutional extremism in the Netherlands*: *A serious threat to the democratic legal order?* (The Hague: Ministry of the Interior and Kingdom Relations, 2023), 6.
134 The Ministers of Justice and Security, of Foreign Affairs, of Economic Affairs and Climate Policy, of Social Affairs and Employment, of the Interior and Kingdom Relations, and the State Secretary for the Interior and Kingdom Relations to the Speaker of the House of Representatives, December 12, 2023, https://www.nctv.nl/documenten/publicaties/2023/12/12/versterkte-aanpak-online-inzake-terroristische-en-extremistische-content.

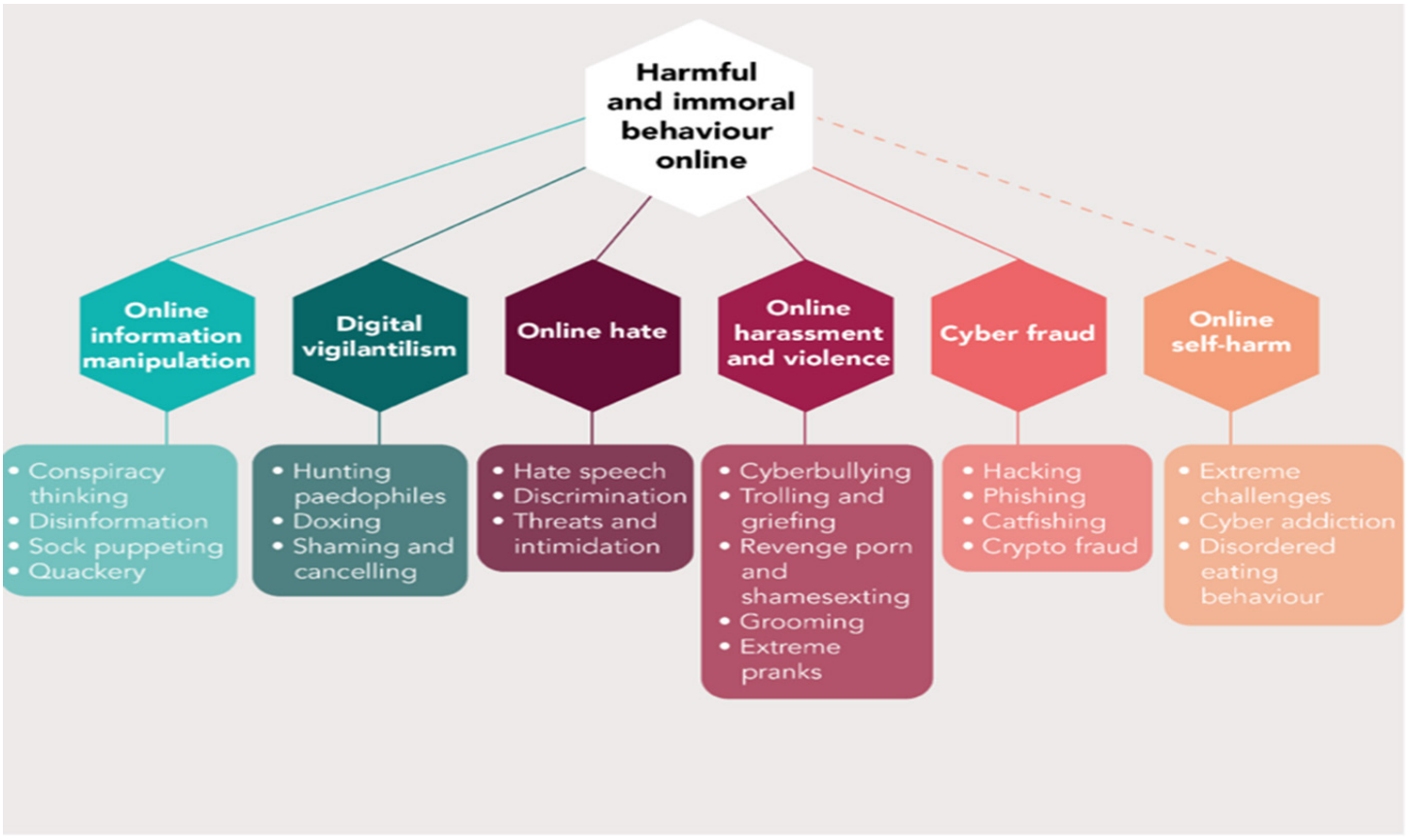
4.2.4 Harmful Content

Both harmful content and disinformation are mentioned in the DSA. While the aim of DSA is to create a safer online environment and impose a range of obligations on different sizes of online service providers, it does not provide a clear distinction between illegal and harmful content, leaving a broad margin of appreciation to the respective online service providers.[135]

However, in February 2025, the European Commission and the European Board for Digital Services endorsed the official integration of the already existing voluntary Code of Practice on Disinformation into the framework of the DSA. The aim of this Code of Conduct is to combat disinformation risks while fully protecting the freedom of expression and improving transparency under the DSA.[136]

Against this background, harmful content can be understood to consist of anything, such as an image, audio, video, or text that could cause serious offence, distress, or harm to an individual, a group of persons, institutions, or is harming the democratic legal order. In July 2021, the Rathenau Institute conducted research into harmful and immoral online behaviour and identified six categories of harmful behaviour that can severely impact individuals, groups, and society as a whole.[137]

Figure 1: 2021 Rathenau Institute Categorisation of Harmful and Immoral Behaviour Online138



135 "In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression (Part 1 of 3)," Britt van den Branden et al., DSA Observatory, published August 2, 2021, https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/#:~:text=In%20Article%202(g)%20DSA,tools%20used%20in%20content%20moderation.

136 "The Code of Conduct on Disinformation," European Commission, published February 13, 2025, https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation.

137 Mariëtte van Huijstee et al., *Online ontspoord – Een verkenning van schadelijk en immoreel gedrag op het internet in Nederland* (The Hague: Rathenau Instituut, 2021).

138 Van Huijstee et al., *Online ontspoord*.

The determination of the underlying harm is inherently subjective and heavily reliant on the broader context surrounding the content in question.[139] Under the UK's Online Safety Act 2023,[140] harm is narrowly defined as physical and psychological harm to individuals resulting from specific illegal activities. In contrast, the EU, through the DSA, adopts a broader perspective, encompassing harm to both individuals and society and considers harm as stemming from systemic issues rather than isolated individual behaviours.[141] These divergent approaches to the concept of harm indicate that content moderation decisions related to so-called harmful content may differ significantly depending on how harm is defined in each jurisdiction. Furthermore, the absence of clear and precise definitions concerning the specific content to be moderated, the methods of moderation, as well as data retention practices, may pose significant challenges to HSPs in ensuring the effective moderation of such content.[142]

It is thus relevant to highlight that in the context of this present research, not all harmful online content is unlawful or constitutes a criminal offence. For example, some extreme online challenges, such as the red pepper challenge, which encourages individuals to film themselves while eating very spicy peppers, are deemed lawful. However, the choking challenge, which encourages teenagers to choke themselves just before passing out and posting it online, is deemed illegal and has led to the arrest of four teenagers, as this encouragement could constitute a criminal offence.[143]

Another example of harmful online content is disinformation. According to the Dutch government, one of the forms of disinformation refers to intentionally spreading misleading information, often with malicious intent. Disinformation can have a negative impact on national security.[144] The underlying purpose of disinformation is always to deceive people or cause harm.[145] While disinformation is harmful, it can be lawful and sometimes unlawful or illegal (see paragraph 4.2.1 on the distinction). Spreading disinformation about the COVID-19 virus can be harmful as it may lead to fewer vaccinations, but when it incites violence against scientists and doctors, it becomes illegal. Spreading disinformation about educational programmes related to puberty and sexuality is not only considered to be harmful but can also be considered unlawful. In a recent case, the Christian organisation 'Civitas' publicly linked the 'Rutger Foundation', an NGO that had developed such a sexual educational programme for children, to pedophilia without any factual basis. The Rutger Foundation accused Civitas of defamation, slander and incitement to hate and discrimination. The court ruled in this case that the interests of the Rutger Foundation of not being exposed to serious allegations outweighs Civitas' freedom of expression regarding - in their view - overly sexualised education.[146]

Harmful online content can also be lawful, such as most – but not all – parody, satire, artistic content or legitimate political speech that is protected under the freedom of speech. The use of cartoons can lead to difficult cases in assessing the lawfulness.[147] An example is the case

139 This was also highlighted by experts during the Expert Roundtable on 15 May 2025 as well as experts interviewed for this study, see e.g. interview conducted on 11 June 2025, on file with research team.
140 Online Safety Act 2023, 2023 c. 50 UK Public General Acts (2023), https://www.legislation.gov.uk/ukpga/2023/50/contents.
141 Benjamin Farrand, "How do we understand online harms? The impact of conceptual divides on regulatory divergence between the Online Safety Act and Digital Services Act," *Journal of Media Law* 16, no. 2 (2024): 12-15, https://doi.org/10.1080/17577632.2024.2357463.
142 Adam Juszczak, and Elisa Sason, "Recalibrating Data Retention in the EU," *Eucrim*, no. 4 (2021): 238-266, https://eucrim.eu/articles/recalibrating-data-retention-in-the-eu/.
143 "Aanhoudingen voor 'choking challenge'," Nationale Politie, accessed September 10, 2025, https://www.politie.nl/nieuws/2025/januari/28/07-aanhoudingen-voor-choking-challenge.html.
144 "Desinformatie," NCTV, accessed September 10, 2025, https://www.nctv.nl/onderwerpen/desinformatie.
145 The Minister of the Interior and Kingdom Relations, and the State Secretary for the Interior and Kingdom Relations to the Speaker of the House of Representatives, 1, https://zoek.officielebekendmakingen.nl/kst-30821-173.pdf.
146 Civitas has the right to disagree with the content of RutgerFoundation's educational programme, and to express that opinion. However, this right to freedom of expression (and freedom of religion) is not unlimited. The serious accusations directed at Rutger Foundation - even if framed as opinions - must be supported by some factual basis. This applies all the more because many of the accusations by Civitas were presented as factual claims. Although Civitas refers to specific parts of Rutger Foundation's teaching materials to support its views, as outlined above, it presents an incomplete, incorrect, and misleading representation of those materials. These references cannot substantiate the accusations made by Civitas. The court therefore rules that the accusations made by Civitas are unlawful toward the Rutger Foundation. \ Stichting Rutgers v. Stichting Civitas Christiana, C/16/590766 / KG ZA 25-121 Rechtbank Midden-Nederland (2025), https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBMNE:2025:1778.
147 David Keane, "Cartoon Violence and Freedom of Expression", *Human Rights Quarterly 30 (2008),* pp. 845-875, https://www.jstor.org/stable/20486714.

regarding a cartoon that implies that Jews invented the Holocaust.[148] A Dutch court of first instance acknowledged that the cartoon, taken by itself, is distasteful and commonly constitutes both a criminal offence under Article 137c of the Dutch Penal Code and an offence towards Jews. It concluded that due to the context it cannot be treated as illegal content, and thus falls under the freedom of expression and is lawful.[149] However, in the Appeals Court and in the Supreme Court, the cartoon was eventually found illegal.[150] These examples illustrate that harmful content can be lawful, unlawful or illegal and thus constitute a criminal offence.

4.2.5 Borderline Content

Most illegal content can be identified more objectively based on the elements of the respective crime, whilst the question of what constitutes borderline content is more subjective. There is no universally accepted definition of borderline content in the context of violent extremism and terrorism. The common phrase 'awful but lawful' or 'legal but harmful' does not fully grasp the issues pertaining to borderline content. What constitutes borderline content is more subjective and depends on the type of content, the user, the audience, and the context. Considering the generally harmful nature of borderline content, it would be helpful to be able to recognise borderline content more objectively to allow policy makers, practitioners, and online platforms to develop appropriate and consistent policies and responses to address borderline content whilst respecting the freedom of speech.

In the context of the EU Radicalisation Awareness Network (RAN), borderline content in the context of prevention and countering violence is described as "content that is not explicitly considered illegal but may still promote extremist narratives, incite violence, or have a polarising effect on society as a whole."[151]

At a national level, the Dutch Enhanced Approach Online for Extremist and Terrorist Content Enhanced Approach refers to borderline content as 'legal yet harmful.' The starting point is that this content is initially legal. Yet it warns: *"[...]* there is an increasing presence of (legal) online content that does not have a strictly terrorist character and does not always lead to violence, but still undermines the safety of citizens and institutions, for example, by inciting hatred or agitation or because the content normalises extremist ideologies.*"*[152]

Borderline content is often difficult to detect because of the use of humour, irony, or memes and can thus constitute *implicit* extremist content (see more elaborate chapter 5). Furthermore, the perception of what constitutes borderline content and consequently should be moderated differs between practitioners, policymakers, and researchers on one hand and online service providers on the other hand.[153] For example, Meta moderates so-called borderline content on its platforms, defining it as "types of content that are not prohibited by [Meta's] Community Standards but that come close to the lines drawn by those policies." Meta justifies moderating such content when related to nudity and sexual activities, violent and graphic content, bullying, harassment, and hate speech, and selling of prohibited goods by saying that such content albeit not violating any guidelines is "sensationalist or provocative and can bring down the overall quality of discourse and that users have frequently told [Meta] that they do not like encountering these forms of content."

148 Openbaar Ministerie v. Arabisch Europese Liga Nederland, 16/610301-09 [P] Rechtbank Utrecht (2010), https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBUTR:2010:BM1984&showbutton=true&keyword=spotprent&idx=13.
149 OM v. AEL, 16/610301-09 [P].
150 Arrest Hoge Raad, 27 March 2012, https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:PHR:2012:BV5623.
151 "Call for participants: RAN C&N meeting on How to deal with borderline content (related to hate speech, meme culture, humour etc.) from the perspective of public trust?" European Commission, published March 1, 2024, https://home-affairs.ec.europa.eu/news/call-participants-ran-cn-meeting-how-deal-borderline-content-related-hate-speech-meme-culture-humour-2024-03-01_en#:~:text=Borderline%20content%20in%20the%20context,on%20society%20as%20a%20whole.
152 The Ministers of Justice and Security, of Foreign Affairs, of Economic Affairs and Climate Policy, of Social Affairs and Employment, of the Interior and Kingdom Relations, and the State Secretary for the Interior and Kingdom Relations to the Speaker of the House of Representatives.
153 "Content borderline to the Community Standards," Meta, accessed April 30, 2025, https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards/.

It is crucial to provide a common understanding of what constitutes borderline content in order to develop rule-of-law based moderation policies and practices. Yet, the nature of content must always be seen in the individual context, meaning that a piece of content could constitute borderline in one specific context, while in another it is of no concern.[154] For example, the European Court of Human Rights (ECtHR) in *Leroy v. France* determined that the publication of a caricature of the 9/11 terrorist attacks, accompanied by a caption inspired by an advertisement, was unlawful given the context of those recent attacks.[155] The defining characteristic of borderline content is that, while it may not always constitute illegal content nor a breach of the ToS of online platforms, it may nonetheless be harmful. Such content may facilitate the spread of terrorist and extremist ideologies, thereby complicating efforts to effectively moderate such content.[156] It is important that a rule of law-based definition does not rely on ToS of private actors but rather existing legal frameworks (national and international/EU).

In response to these concerns, in the context of the European Counter Terrorism Centre (ECTC) Advisory Network Conference in 2023, experts suggested establishing a comprehensive framework including a list of content types deemed borderline, alongside mechanisms to ensure independent oversight and multistakeholder consultations.[157] Also in 2023, the European Commission developed a handbook on borderline content to guide tech companies in the identification of harmful but legal content that can lead towards radicalisation.[158] The Global Internet Forum to Counter Terrorism (GIFCT) contributed to the development of this handbook through a dedicated paper by online platforms, in particular on how to identify and limit the spread of borderline content that can lead to violent extremism and terrorism. In its paper, GIFCT does not provide a definition but acknowledges that "the term 'borderline content' is by its nature subjective, and most often used to denote a range of online policy or content areas that have overlap with terrorist content."[159] At the same time, GIFCT recognises the need to define the 'parameters' of borderline content.

According to GIFCT, borderline content in the context of violent extremism and terrorism can be found in different categories of content, such as:[160]

- Hate speech;
- Incitement to violence;
- Violent content/graphic;
- Weapons and instruction materials;
- Misinformation;

As seen from the above, one of the challenges of defining borderline content is that it is an umbrella term that can encompass many different types of content, ranging from self-harm to disinformation to gore content.[161] The focus of this research is on how to identify and address borderline, defined as implicit extremist content that can lead to radicalisation towards extremism and terrorism.

154 Saltman, and Hunt, "Borderline Content," 3-4.
155 Leroy v. France, 36109/03 European Court of Human Rights § 5 (2008), https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-88657%22]}.
156 Stuart Macdonald, and Katy Vaughan, "Moderating borderline content while respecting fundamental values," *ECTC Advisory Network Conference* (2023): 2-3, https://www.europol.europa.eu/cms/sites/default/files/documents/macdonald_vaughan.pdf.
157 Ibid.
158 EU Internet Forum, *EU Internet Forum at 10 YEARS: Celebrating the achievements of the first decade's cooperation to fight harmful and illegal content online* (Brussels: European Commission, 2024), https://home-affairs.ec.europa.eu/document/download/0fb0be8a-c145-4948-a260-0a6be11ffc53_en?filename=EU%20Internet%20Forum%20Brochure.pdf&prefLang=uk.
159 Saltman, and Hunt, "Borderline Content."
160 Saltman, and Hunt, "Borderline Content," 6-7.
161 Macdonald, and Vaughan, "Moderating borderline content while respecting fundamental values," 4-5.

4.2.6 Hate Speech

As hate speech has been identified by online platforms and GIFCT as one of the most common categories of borderline content in relation to extremism and terrorism,[162] this terminology requires a more thorough assessment. First of all, it must be noted that there is no universally accepted definition of hate speech, and the term extends the scope of incitement to discrimination, hostility or violence, which is defined under international human rights law and will be addressed in the following section (Annex 4).

Hate speech – whether committed offline or online – has an impact on several human rights and can notably infringe on the right to religious beliefs and the right to be free from discrimination. Additionally, measures countering hate speech can, in turn, also lead to violations of certain fundamental rights, notably the right to freedom of expression, and the right to freedom of association and peaceful assembly.

The United Nations (UN),[163] the EU[164] and the Council of Europe[165] have adopted working definitions of hate speech, and many countries have criminalised hate speech. In addition, several of the bigger platforms have also included a definition of hate speech in their terms of reference, policies or community guidelines.

At the EU level, the DSA specifically refers to hate speech as a form of illegal content that falls in the first category of systematic risks that should be assessed by VLOPs and very large online search engines (VLOSEs). The DSA also refers to a Code of Conduct on Countering Illegal Hate Speech Online, which was revised in early 2025 and is now integrated into the DSA.[166] This Code refers to:

> *"illegal hate speech as defined by applicable laws, including the Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, as transposed in national jurisdictions, as well as possible forthcoming updates to this Framework Decision, where relevant."*[167]

Once again, there is no uniformity in the way hate speech is defined. The risk of having different definitions – especially those that are defined too broadly or too narrowly can have an impact on what content is removed online. The risk of content being:

- Too broad: if the definitions that are being applied are too broad, this means that more content that is deemed harmful but is strictly lawful is being removed, which could constitute a violation of the right to freedom of expression and other human rights;
- Too narrow: if the definitions that are being applied are too narrow, this means that harmful content that is illegal or legal is not being removed and could lead to violent extremism or terrorism.

In fact, many HSPs provide for a broad definition of hate speech in their community guidelines or ToS, which can entail many different forms of unwanted content, ranging from terrorist conduct to other illegal content or borderline content. For example, several HSPs have signed the Code of Conduct on Countering Illegal Hate Speech Online + in January 2025. However, this

162 Saltman, and Hunt, "Borderline Content," 6-7.
163 UN Office on Genocide Prevention and the Responsibility to Protect, *United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences* (New York: UN, 2020), 10, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf.
164 European Commission, *Code of Conduct on Countering Illegal Hate Speech Online* (Brussels: European Commission, 2016), https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_en.
165 Committee of Ministers to member states on "hate speech," CM/Rec (2022)16 – recommendation of the Committee of Ministers to member States on combating hate speech, 20 May 2022, Council of Europe, https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680a67955%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}.
166 "The Code of conduct on countering illegal hate speech online +," European Commission.
167 European Commission, *Code of Conduct on Countering Illegal Hate Speech Online +* (Brussels: European Commission, 2025), fn. 4, https://ec.europa.eu/newsroom/dae/redirection/document/111777.

code of conduct only lays out broad policies in relation to appeal mechanisms, cross-platform cooperation, and transparency for hate speech, without providing a uniform definition of the term.[168] In fact, studies have shown that major platforms, including Instagram, Reddit, and TikTok, have expanded their definition of hate speech over time to include ever more characteristics and types of content.[169]

4.3 Towards a Working Definition: Implicit Extremist Content that is Harmful

Considering the different uses and definitions of the term borderline content, for the purpose of this study, the research team decided to refrain from using the term borderline and instead refers to **implicit extremist content that is harmful** (see paragraphs 1.3 and 1.4)**.** Moreover, this study will use the definitions laid out below of other critical terms relevant to the detection and moderation of content that can lead to radicalisation towards extremism and terrorism:

'Extremist' refers to the fact that the content disseminates exclusionary and hateful narratives that may contribute to radicalisation toward 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 persons, institutions or to the democratic legal order, and that is not protected under international human rights law.

Once implicit extremist content is assessed – allowing to lift the veil and uncover the true meaning that was intended to be concealed – the content could be identified as 'unlawful' (including content that could potentially qualify as terrorist or other forms of illegal content) and/or 'lawful but harmful' extremist content. Unlawful could refer to unlawful and/or illegal.

4.4 Definitions by Platforms

Based on the selection of the three selected platforms as described in chapter 3, this chapter assesses how these platforms define illegal, terrorist, and borderline content in their Terms of Use (ToU) and/or Community Guidelines (CG) which form the basis for many of their detection and moderation policies and ultimately practices in relation to implicit extremist content that is harmful.

4.4.1 Instagram

Instagram has a ToU, Community Standards and a range of policies that regulate illegal, terrorist, and borderline content. The Community Standards of Instagram determine what content is and is not allowed on Instagram. These Standards have been developed on the basis of feedback received from people and experts in the field of technology, public safety, and human rights. However, it is not clear whether 'people' refers to users of Instagram and how (frequently) the experts have been consulted to provide feedback on these standards.

Terrorist content

Instagram refers directly to the TCO and explains how competent authorities can report content that violates the TCO and thus relies on their definition, stating that terrorist content, in the form of texts, images, sound recordings or videos, including live transmission, is content that:

168 European Commission, *Code of Conduct on Countering Illegal Hate Speech Online +*.
169 Jacob Mchangama et al., *Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies* (Nashville: The Future of Free Speech, 2023), 104-105, https://futurefreespeech.org/wp-content/uploads/2023/07/Community-Guidelines-Report_Latest-Version_Formated-002.pdf.

- Incites others to commit terrorist offences, such as by glorifying terrorist acts, by advocating the commission of such offences;
- Solicits others to commit or to contribute to the commission of terrorist offences;
- Provides instruction on the making or use of explosives, firearms or other weapons, or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of terrorist offences; or
- Constitutes a threat to commit a terrorist offence.

Under the Community Standards and Policies, Instagram has developed an approach on how to deal with organisations or individuals that incite or engage in violence. Subsequently, two categories on the different seriousness of violence have been developed. The first category includes: hate organisations, criminal organisations – including those designated by the United States (US) government as specially designated narcotics trafficking kingpins (SDNTKs) – and terrorist organisations – including entities and individuals designated by the US government as foreign terrorist organisations (FTOs) or specially designated global terrorists (SDGTs). In doing so, Instagram defines terrorist organisations and individuals as a non-state actors that:

- Engage in, advocate or lend substantial support to purposive and planned acts of violence,
- Cause or attempt to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians
- Act with the intent to coerce, intimidate and/or influence a civilian population, government or international organisation in order to achieve a political, religious or ideological aim.

Ultimately, Instagram's Community Standards and Policies state that any forms of glorification, support or representations of these organisations will lead to the removal of related content.
Borderline content – content that comes close to lines drawn by policies

Instagram identifies borderline content, but is not limited to the following categories:

- Adult nudity and sexual activity,
- Violence and graphic content,
- Bullying and harassment, hate speech, violence and incitement.

The underlying reasoning is that this content can be considered to be "sensationalist or provocative and can bring down the overall quality of discourse on our platform, especially because people have frequently told us that they do not like encountering these forms of content." [170]

The way borderline content is described is very broad and subjective. According to the notion that certain types of content are being removed based on user feedback and the preferences of what users want to see on the platform, one could assume that if a person with conservative views reviews certain content on Instagram and considers the content a violation, it will be removed based on that feedback.

Content likely to violate community standards

Instagram goes one step further and also identifies content that could likely violate Community Standards. The following categories of content that could likely violate community standards are used by Instagram:

- Hate speech: see below
- Violence and incitement: this includes threats to violence that could result in death, serious injury, but also soliciting others to commit violence, instructions on how to make or use

170 "Content borderline to the Community Standards," Meta.

weapons, or explosives with the intent to cause harm, or threats of violence related to voting.
- Suicide, self-injury and eating disorders;
- Bullying and harassment;
- Graphic violence;
- Adult nudity and sexual activity;
- Posts buying, selling, trading or promoting restricted goods or services;
- Spam, fake accounts and scams.

According to Instagram, it aims to minimise possible harm to its community and thus takes steps to reduce exposure to such content that is likely to violate its community standards. [171]

Hateful conduct

Instagram's Community Standards have a policy in place to deal with hateful conduct and remove dehumanising speech, allegations of serious immorality or criminality, and slurs. It defines hateful conduct (*see below*) and indicates what kind of content may not be posted. It defines two categories of hateful conduct that may not be posted.

The first category concerns content targeting a person or group of people on the basis of their protected characteristic(s) or immigration status in written or visual form through:

- Dehumanising speech in the form of comparisons to or generalisations about animals, pathogens or other sub-human life forms;
- Allegations of serious immorality and criminality;
- Calls and hopes for the following harms;
- Harmful stereotypes historically linked to intimidation or violence,
- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image;
- Mocking people for having or experiencing a disease;
- Content that describes or negatively targets people with slurs. Slurs are defined as words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression and violence.[172]

The second category of hateful content that is prohibited targets a person or group of people on the basis of their protected characteristic(s) in written or visual form through:

- Calls or support for exclusion or segregation or statements of intent to exclude or segregate;
- Insults;
- Expressions that suggest the target causes sickness;
- Targeted cursing, except certain gender-based cursing in a romantic break-up context.[173]

Instagram provides examples of what is considered a violation of the Community Standards.

4.4.2 Reddit

Reddit refers in the User Agreement to their rules, other terms, and policies with respect to content. According to Reddit rule 1, "everyone has the right to use Reddit free of harassment, bullying, and threats of violence", and those who "incite violence or that promote hate based on identity of vulnerability will be banned". This rule contains two relevant aspects. Firstly, it

171 "Content likely violating our Community Standards," Meta, accessed September 10, 2025, https://transparency.meta.com/fr-fr/features/approach-to-ranking/content-distribution-guidelines/content-likely-violating-our-community-standards#.
172 "Hateful Conduct," Meta, accessed September 10, 2025, https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/.
173 "Hateful Conduct," Meta.

addresses posts relating to threats of violence. Secondly, it addresses hate speech. According to the Reddit Rules, posts that encourage, glorify, incite, or call for violence or physical harm against an individual (including oneself) or a group of people are not permitted. Examples include:

- Posts or comments with a credible threat of violence against an individual or group of people.
- Posts containing mass killer manifestos or imagery of their violence.
- Terrorist content, including propaganda.

Furthermore, the Reddit Rules also prohibit content that glorifies, incites or calls for violence or physical harm, including content that promotes or supports the activities of terrorists or designated terrorist organisations. This includes propaganda material posted by terrorists or designated terrorist organisations and their supporters, expressions of affiliation or support for terrorists or designated terrorist organisations, and glorification of terrorist acts. It also includes content that solicits or incites a person or group to participate, commit, or contribute to terrorist activities.[174] It appears that Reddit does not provide definitions of the terms (designated) terrorist groups or terrorist activities. Borderline content is not mentioned in the Reddit Rules.

Hateful content

Moreover, the Reddit Rules also mention hate speech and do not allow posts against marginalised or vulnerable groups, which include, but are not limited to, groups based on their actual and perceived race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These protected groups also include victims of major violent events and their families.

4.4.3 TikTok

In their ToS and CG, TikTok lays out how to deal with illegal, terrorist, and child sexual abuse material (CSAM) content. Borderline content or content that is at the edge of violating these ToS is not mentioned in particular.[175]

In section 4.5 of the ToU, TikTok prohibits posting, live streaming or sharing any content which:

- Infringes anyone else's rights (such as intellectual property, privacy and/or personality rights of living or deceased people);
- Constitutes, encourages or provides instructions for a criminal offence, or dangerous activities that may lead to serious injury or death or self-harm;
- Spreads harmful misinformation, such as misinformation that incites hate or prejudice or that misleads about or improperly influences elections or other civic processes;
- Contains a threat of any kind, which intimidates or harasses others, including posting any material that is intended to mock, humiliate, embarrass, intimidate, or hurt an individual;
- Is obscene, pornographic or promotes sexually explicit material (e.g. by linking to adult or pornographic websites);
- Is hateful or inflammatory;
- Contains or promotes violence or discrimination based on race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender identity, serious disease, disability, immigration status or age;
- Otherwise contains harmful content (such as content that causes physical, mental or moral detriment to minors).

The TikTok CG further provide for three categories of problematic content that are relevant in the context of this research and recognises that online violence can lead to violence offline. Firstly,

174 "How does Reddit fight the dissemination of terrorist content?" Reddit, accessed September 10, 2025, https://support.reddithelp.com/hc/en-us/articles/19003525756564-How-does-Reddit-fight-the-dissemination-of-terrorist-content.
175 "Terms of Service," TikTok, accessed September 15, 2025, https://www.tiktok.com/legal/page/eea/terms-of-service/en.

TikTok does not "allow threats, glorifying violence, or promoting crimes that could harm people, animals, or property." It mentions that it does not allow content that is:

- Threatening or expressing a desire to physically harm people
- Promoting or glorifying violence, such as calling for violent attacks or praising acts of violence
- Promoting theft or the destruction of property
- Sharing instructions on how to commit crimes that may harm people, animals, or property.

Secondly, TikTok addresses hate speech and hateful behaviour. Hate speech and hateful behaviour are understood as attacking, threatening, dehumanising or degrading an individual or group based on their protected attributes. Protected attributes mean personal characteristics that you are either born with, are immutable, or that would cause severe psychological harm if you were forced to change them or were attacked because of them. These include race, ethnicity, national origin, tribe, caste, sexual orientation, sex, gender, gender identity, serious disease, disability, and immigration status. In addition, age and other protected attributes could be considered hateful depending on the context. The Guidelines provide a non-exhaustive list of examples of what behaviour is not permitted.

Furthermore, TikTok also identifies a third category consisting of hateful organisations and persons that spread beliefs or propaganda that encourage violence or hate. Promoting or providing material support to them is not permitted. TikTok refers to violent extremist entities, violent criminal organisations, violent political organisations, hateful organisations, and individuals who cause mass or serial violence. In this category fall:[176]

- Violent extremist entities, understood as non-state groups or individuals who engage in or advocate for violence against civilians, to advance political, religious, ethnic, or ideological objectives, in ways recognised as violating international norms.
- Violent criminal organisations, understood as transnational, national, or local groups that commit serious crimes, including violence, trafficking, and kidnapping.
- Violent political organisations, understood as non-state actors that commit violent acts primarily against state actors (such as a national military) rather than civilians, as part of ongoing political disputes (such as territorial claims).
- Hateful organisations, understood as groups that target people based on protected attributes, dehumanise others, and promote hateful ideologies.

It is not permitted to promote (including any praise, celebration, or sharing of manifestos) or provide material support to hateful organisations, individuals who cause serial or mass violence, or promote hateful ideologies, violent criminal organisations and violent extremists.

4.5 Preliminary Observations, Findings, and Gaps

The lack of clear internationally accepted definitions of terrorism, violent extremism, and hate speech poses a challenge not only for the platforms but also among policy makers and practitioners in determining what constitutes terrorist content, illegal content, extremist content, and borderline content and how to address it in a rule-of-law-based manner. Indeed, research has found that a crucial hurdle in allowing for a human rights compliant content moderation across different platforms is the lack of internationally agreed definitions of other removable content, such as, for example, *hate speech* or *harmful content*.[177] While Article 35 (1) lit. c of the DSA mentions "illegal hate speech" as content that should be moderated in "consideration to the impact of such measures on fundamental rights," there is no universally accepted definition

176 "Safety and Civility," TikTok, accessed September 15, 2025, https://www.tiktok.com/community-guidelines/en-GB/safety-civility?cgversion=2025H2update.
177 ARTICLE 19, *Content moderation and freedom of expression handbook* (London: ARTICLE 19, 2023), 8, https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf.

of what constitutes hate speech. Similarly, the removal of *harmful content* presents a complex challenge, as content that may appear harmful does not always constitute a violation of a platform's ToS and might not always be illegal or unlawful.[178]

While illegal terrorist offences are relatively more clearly defined, the terms (direct and indirect) incitement and glorification of these terrorist offences are not clearly defined. A broad interpretation of these terms can lead to unlawful interference with the right to freedom of expression. The UN Human Rights Committee emphasised that "such offences as 'encouragement of terrorism' and 'extremist activity' as well as offences of 'praising', 'glorifying', or 'justifying' terrorism, should be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression."[179]

One of the challenges in defining borderline content is that it is an umbrella term that can encompass many different types of content, ranging from self-harm to disinformation to gore content. [180]

Platforms refer to borderline content as content that may not (yet) violate the policies of the platforms and is connected to a range of policy areas, making a one-size fits all approach to borderline content difficult. Moreover, platforms encounter significant difficulties in implementing a standardised approach to implicit extremist content due to the divergent norms and values across different communities and cultures.[181] Similarly, implicit extremist content that may be deemed harmless in one context can be perceived as harmful in another, with these contextual interpretations evolving over time. Therefore, it is essential for future policy development and platform practices to address the complexities associated with so-called borderline content, defined for this study as implicit extremist content, ensuring a more effective response to these challenges.

Moreover, the lack of a clear definition of what constitutes borderline content may lead to the inadvertent removal of lawful content, thereby infringing upon the fundamental right to freedom of expression.[182]

178 Farrand, "How do we understand online harms?" 6.
179 General comment No. 34, 102 UN Human Rights Committee (2011), https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.
180 Macdonald, and Vaughan, "Moderating borderline content while respecting fundamental values," 4-5.
181 Arnav Arora et al., "Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go," *ACM Computing Surveys* 56, no. 3 (2023): 9, https://doi.org/10.1145/3603399.
182 Dušan Popović, "The Digital Platforms' Sisyphean Task: Reconciling Content Moderation and Freedom of Expression," in *Repositioning Platforms in Digital Market Law* (Cham: Springer, 2024), 100, https://doi.org/10.1007/978-3-031-69678-7_4.

5. The Use of Online Content for Terrorist and Extremist Purposes

While the previous chapter highlighted the legal frameworks and scope of definitions applicable to online content as well as relevant policy documents of the three online platforms selected for this study, this chapter will focus on the potentially terrorist, illegal and implicit extremist content that is harmful in itself. First, the focus will be on what the content is about, the way it is presented, the impact it has on society, and to what purpose such content is being disseminated online. Examples pulled from the open-source intelligence (OSINT) research will be used to illustrate where relevant. Furthermore, this chapter will elaborate on how advanced technologies are being used to generate, amplify, and disseminate content in a way that increases both the reach of and engagement with such content.

With a view to developing a pilot codebook on the detection of implicit extremist content that is harmful, this chapter will also flag relevant characteristics of online content that can function as potential indicators to determine whether the content qualifies as terrorist, illegal or implicit extremist. These indicators will ultimately be included in the pilot codebook introduced in chapter 7 to test the feasibility of developing an assessment framework of online content that leads to radicalisation towards extremism and terrorism.

5.1 Purposes of Online Activities by Extremists

Crises like the COVID-19 pandemic and the Israel-Gaza conflict show how extremist groups exploit turmoil to spread hate speech and fuel polarisation in society.[183] Especially, the rapid spread of online content amplifies hate speech and aids radicalisation, as extremists use disinformation and polarised narratives to recruit and gain support.

For instance, jihadist videos are disseminated across multiple platforms, where they are swiftly downloaded.[184] Offering this easy access content facilitates, intensifies, and accelerates online (self-) radicalisation.[185] However, this is not a new trend. For instance, investigations of the 2010 attack in which a university student assaulted a Member of Parliament in the UK revealed that the perpetrator had accessed websites and audiovisual sermons delivered by radical preachers.[186] One key concern for authorities in relation to extremists' and terrorists' use of the online sphere is its radicalising potential, in particular among younger people.[187] While the process of terrorist radicalisation via the internet is not a novel issue, online radicalisation has gained greater attention in the context of extremist ideologies in recent years.[188]

Beyond that, extremist and terrorist groups exploit online services for recruitment and mobilisation, to fuel societal polarisation, for networking and plotting.[189] The capacity for remote recruitment enables recruiters to expand their reach, allowing them to engage with a broader pool of candidates, assess their compatibility with specific groups through online platforms, and

183 Annelies Pauwels, and Maarten van Alstein, "Polarisation: A short introduction," *RAN Spotlight* (April 2022): 10-11, https://home-affairs.ec.europa.eu/document/download/edf83900-6f86-4ef3-8bce-2d20e2e1d06f_en?filename=ran_spotlight_polarisation_en.pdf.
184 Ali Fisher, "Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence," *Perspectives on Terrorism* 9, no. 3 (2015): 3, https://pt.icct.nl/sites/default/files/import/pdf/2-swarmcast-how-jihadist-networks-maintain-a-persistent-online-presence-by-ali-fisher.pdf.
185 Christina Schori Liang, "Cyber Jihad: Understanding and Countering Islamic State Propaganda," *Geneva Centre for Security Policy* (February 2015): 2, https://www.gcsp.ch/sites/default/files/2024-12/PP2-2015%20-%20LIANG%20-%20Cyber%20Jihad%20-%20Draft%20F.pdf.
186 Liang, "Cyber Jihad," 2.
187 Europol, *EU TE-SAT 2024*, 10; NCTV, *Dreigingsbeeld Terrorisme Nederland: december 2024* (The Hague: Ministry of Justice and Security, 2024), 3, 5 & 18-28, https://www.nctv.nl/onderwerpen/dtn/documenten/publicaties/2025/06/17/dreigingsbeeld-terrorisme-nederland-june-2025.
188 Van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms?* 4-7.
189 Sam Hunter et al., "The Metaverse as a Future Threat Landscape: An Interdisciplinary Perspective," *Perspectives on Terrorism* 18, no. 2 (June 2024): 65, https://pt.icct.nl/sites/default/files/2024-06/Research%20article_Hunter.pdf.

connect with individuals across diverse geographic regions.[190] Notably, following the terrorist attacks on 7 October 2023, there has been a noticeable increase in radicalisation, recruitment, and fundraising activities around the world.[191]

Furthermore, academic research has demonstrated that extremists exploit social media not only for online recruitment but also as a tool to organise offline events, bringing together individuals from the same locality.[192] These events may serve as opportunities for further recruitment, thereby facilitating both virtual and physical networks of influence. The growing concerns with online networks such as Terrorgram Collective and 764 network demonstrate how social media and gaming platforms are used to recruit members and incite to offline violence. [193]

Terrorgram Collective is an online neo-fascist network that spreads online propaganda and incites violence. It has been linked to several plots and attacks, including a shooting at an LGTBTQI+ bar in Slovakia and a plot against energy facilities in the US. Terrorgram consists of different channels operating on Telegram, which has failed to moderate this group. Terrorgram has been designated by the UK and the US as a terrorist organisation. [194]

Additionally, the online sphere is used by terrorist and extremist groups for financing purposes.[195] Such funds are crucial to the operational continuity of these organisations, as they are used not only to finance their activities but also for logistical purposes, including training simulations and the procurement of weapons.[196]

Furthermore, the online sphere is used by terrorists and extremists in plotting and preparing attacks. Terrorist organisations mainly resort to illicit methods to obtain weapons.[197] The widespread accessibility of firearms on the black market, including through online platforms, facilitates this illicit acquisition.[198] Furthermore, research has highlighted the increasing use of privately manufactured firearms by different terrorist and extremist groups and individuals.[199] Notably, right-wing extremists have turned to 3D-printing technology to manufacture firearms. Several individuals from the right-wing violent extremist group Sturmjäger Division were arrested in Croatia in 2023 for circulating 3D-printed weapons manuals online and providing instructions on how to create pipe bombs.[200] Although possession or dissemination of 3D-printing manuals is not criminalised in many jurisdictions, sharing terrorist publications, including instructions on how to make weapons, can be prosecuted. A man from Liverpool has been convicted for sharing a manual filled with instructions on how to build weapons, including shotguns, nail bombs, and explosives.[201]

190 Hunter et al., "The Metaverse as a Future Threat Landscape," 71.
191 Europol, *EU TE-SAT 2024*, 7.
192 Williams et al., "The Online Extremist Ecosystem," 5.
193 Gabriel Weimann et al., "White Jihad: Fused Extremism?" *Terrorism and Political Violence* (2025), https://gnet-research.org/2024/01/19/764-the-intersection-of-terrorism-violent-extremism-and-child-sexual-exploitation/.
194 "Beyond the Collective: Understanding Terrorgram's efforts to infiltrate the mainstream on Telegram," Steven Rai, Institute for Strategic Dialogue, published August 24, 2024, https://www.isdglobal.org/digital_dispatches/beyond-the-collective-understanding-terrorgrams-efforts-to-infiltrate-the-mainstream-on-telegram/; Colin Clarke et al., "Why the Terrorgram Collective Designation Matters," *Lawfare* (2025), https://www.lawfaremedia.org/article/why-the-terrorgram-collective-designation-matters.
195 Hunter et al., "The Metaverse as a Future Threat Landscape," 65.
196 Hunter et al., "The Metaverse as a Future Threat Landscape," 68.
197 Annelies Pauwels, and Merlina Herbach, "Buy It, Steal It, Print It: How Right-Wing Extremists In Europe Acquire Firearms And What To Do About It," *ICCT* (December 2024): 2, https://icct.nl/publication/buy-it-steal-it-print-it-how-right-wing-extremists-europe-acquire-firearms-and-what-do.
198 The European Commission to the European Parliament and Council of Europe, February 12, 2015, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52015DC0624.
199 Pauwels, and Herbach, "Buy It, Steal It, Print It," 11.
200 Europol, *EU TE-SAT 2024*, 38.
201 He has been convicted for preparation of terrorist acts, dissemination of terrorist publications and possession of material likely to be used for a terrorist attack. "Young man from Liverpool convicted of preparing for acts of terrorism," *Greater Manchester Police*, published February 23, 2024, https://www.gmp.police.uk/news/greater-manchester/news/news/2024/february/young-man-from-liverpool-convicted-of-preparing-for-acts-of-terrorism/.

5.1.1 Polarisation

Polarisation describes the process of views and opinions becoming increasingly opposed and sharpened. While polarisation itself is not inherently a negative phenomenon – it can after all also drive societal change, it can contribute to the escalation of tensions with serious impacts on democratic systems.[202] Notably, polarisation becomes undesired in societies if it creates persistent antagonisation and prevents different groups from constructive communication with each other, leading to the destruction of normal social interaction.[203] If the opposing views of groups vis-à-vis the other grow deeper, and also have an emotional impact on how one perceives the one's own group and the other group, the polarisation is considered problematic and harmful.

The rapid proliferation of online content further intensifies the dissemination of hate speech and facilitates the radicalisation of individuals by extremist and terrorist actors, preventing constructive dialogue with other groups.[204] A study underscores that passive exposure to such narratives online was one of the most perilous aspects of the internet.[205] This not only increases the creation of online communities but also reinforces radical narratives where individuals with similar views can exchange and legitimise extreme perspectives without engaging in constructive dialogue with others outside this group.[206] Additionally, the anonymity in the online sphere facilitates the expression of behaviours and attitudes that would otherwise be deemed socially unacceptable.[207]

The OSINT research conducted for this study confirms these trends of the establishment of certain echo chambers. A post sharing a headline by the Dutch Newspaper 'Telegraaf', quoting a Dutch politician and his assessment of the Amsterdam riots was shared in two different subreddits. In the subreddit r/FreeDutch that catered to like-minded people, the post received 25 likes, whereas the same post shared on the subreddit r/Politiek received no likes. Respective groups not only leverage social media platforms to disseminate their ideologies but also to foster a collective group identity. The same user who shared this post almost exclusively 'debates' in spaces which are already more aligned with his extremist positions. Another post in which the user links to a biased news report and uses the cover image of young non-white people running through the street and an ambiguous title ("Dit is dus wat je krijgt", translated as "This is what you get". to infer something that can only be truly understood by those already primed for far-right content, received 32 likes. The user also posts this in r/Nederland, where the post receives no votes and less positive engagement. Hence, it appears that this user goes to the more permissive subreddits to post explicit, illegal and implicit extremist content. Additionally, far-right users engage with each other's posts and use similar arguments to boost these views.

Beyond fostering the collective group identity, these examples also point to a perceived need for retaliation or self-defence, implicitly triggering actions that are necessary to protect the group from reputational damage, or instilling the idea of a compensation or reward for undertaking action, and exerting peer pressure, or referencing an implicit code of honour.

The COVID-19 pandemic further increased online activity during social isolation and helped spread rigid, black-and-white worldviews, encouraging group conformity and hostility toward out-

202 Pauwels, and Van Alstein, "Polarisation," 10-11.
203 Ron van Wonderen et al., *Theorieën en aanpakken van polarisatie — Verkorte rapportage van bevindingen en conclusies* (Utrecht: Kennisplatform Integratie & Samenleving, 2020), 5, https://www.kis.nl/sites/default/files/2022-06/theorieen-en-aanpakken-van-polarisatie-samenvatting.pdf.
204 Annelies Pauwels, and Will Baldet, "The Role of Hotbeds of Radicalisation," *RAN Spotlight* (April 2022): 16, https://home-affairs.ec.europa.eu/document/download/edf83900-6f86-4ef3-8bce-2d20e2e1d06f_en?filename=ran_spotlight_polarisation_en.pdf.
205 Guri Mølmen, and Jacob Ravndal, "Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists," *Behavioral Sciences of Terrorism and Political* Aggression 15, no. 4 (2023): 268, https://doi.org/10.1080/19434472.2021.1993302.
206 Mølmen, and Ravndal, "Mechanisms of online radicalisation," 268.
207 Mølmen, and Ravndal, "Mechanisms of online radicalisation," 268.

groups.[208] For instance, thousands of Australian and Canadian right-wing groups on Facebook shared similar themes on violence, national identity, and racial references that were influenced by their socio-historical contexts.[209] Protest movements driven by anti-establishment rhetoric, such as distrust in the government, also emerged in several countries, using COVID-19 restrictions to fuel anti-institutional discourse. The pandemic period thus served as a catalyst for right-wing extremist groups to spread disinformation and conspiracy theories in order to promote anti-democratic and anti-institutional narratives and agendas.[210] These developments also highlight a specific indicator relevant to recognising the presumed intent to cause harm through the shared content, namely the motivation to normalise hateful or violent narratives, and/or to propagate hateful/violent conspiracies.

The OSINT research conducted for this study highlights a prominent approach that couples anti-institutionalism rhetoric with current conspiracy theories, for example, related to COVID-19, in the form of memes that display prominent political figures and carry conspiratorial narratives, thus suggesting the involvement of these individuals in the perceived wrongdoing.[211]

The 2015 terrorist attacks in Paris have generated a significantly polarised discourse concerning Islam and migration.[212] Far-right groups have positioned themselves as defenders of traditional values against perceived threats of multiculturalism and so-called Islamisation by associating security concerns with Islam.[213] This narrative has propagated fear, shaped public attitudes toward immigration, and influenced the rise in support for anti-immigration political parties.[214] The OSINT research conducted for this study also shows that many users associate Islam with violence against Jews following the Amsterdam Riots, often making the argument that Islam, migration from Islamic countries, and left-wing support thereof are incompatible with Dutch values of religious freedom.[215] Some users even argued that banning or removing Islam would not be against religious freedom, as they consider Islam to be an 'ideology'.[216] In addition to being another example of normalising hateful violent narratives and examples of how to instil a perceived need for retaliation or self-defence, these posts are also an example of another relevant indicator, namely, to foster hate or hostility towards an out-group.

In the aftermath of the 7 October attack, there has been significant jihadist propaganda disseminated online, with a clear focus on the responses of both parties to the conflict in Gaza.[217] For instance, al-Qaeda affiliates praised the Hamas-led attack and urged the Muslim community to unite in support of Hamas and the Palestinians. The Islamic State (ISIS) framed the conflict in Gaza as a religious war and circulated publications encouraging Muslims to target synagogues and Jewish economic interests globally.[218] By doing this, an implicit action trigger is included in the narratives, namely, providing a sacred justification for action. Such narratives are also an example of a harmful alliance or affiliation with an extremist or terrorist group, as well as a glorification of crimes committed.

208 Tina Askanius et al., "Far-right extremist narratives in Canadian and Swedish COVID-19 protests: a comparative case study of the Freedom Movement and Freedom Convoy," *Behavioral Sciences of Terrorism and Political Aggression* 17, no. 2 (2024): 2, https://doi.org/10.1080/19434472.2024.2340492.
209 Jade Hutchinson et al., "Mobilizing extremism online: comparing Australian and Canadian right-wing extremist groups on Facebook," *Behavioral Sciences of Terrorism and Political Aggression* 15, no. 2 (2023): 222, https://doi.org/10.1080/19434472.2021.1903064.
210 Askanius et al., "Far-right extremist narratives in Canadian and Swedish COVID-19 protests," 4.
211 See for instance posts #010, #036, and #048 of the OSINT research on file with research team.
212 Teresa Völker, "How terrorist attacks distort public debates: a comparative study of right-wing and Islamist extremism," *Journal of European Public Policy* 31, no. 11, (2023): 3493, https://doi.org/10.1080/13501763.2023.2269194.
213 Völker, "How terrorist attacks distort public debates," 3493.
214 Völker, "How terrorist attacks distort public debates," 3493.
215 See posts #185, #213, #217, #218, #233, #241, #242, #245, #249, and #251 of the OSINT research on file with research team.
216 See posts #219, #220, and #223 of the OSINT research on file with research team.
217 Europol, *EU TE-SAT 2024*, 28.
218 Europol, *EU TE-SAT 2024*, 28.

In the Netherlands, hatred against both Jews and Muslims has increased since 7 October.[219] Building on observations from the OSINT research, it can be seen that several posts by Islamist-inspired users argue that their followers should not look towards the West for solutions to the Israel-Gaza war, but that this solution must come from Muslim countries and preferably from a unified Islamic leadership. Whilst most of their criticism of the West's approach towards Gaza is rooted in historical and current developments and as such falls under legitimate speech, their proposed solutions signal a longing for an Islamic (IS-like) caliphate, which would be problematic, instead of a serious attempt to solve the underlying issues. These posts often reiterate anti-colonialist and anti-nationalist arguments regarding Israel and the region more broadly, coupled with more explicit arguments about the religious importance of Gaza, to make a larger point about these users' Islamist vision for the region, which they then posit as the only way forward.[220]

- **Indicators of implicit extremist content**
  - Fostering a group identity
  - Harmful alliance or affiliation with an extremist or terrorist group.
  - Problematic reference to historical or current context by glorifying or giving a positive portrayal of an event, acts or individuals involved in committing crimes.
  - Presumed intent to cause harm with the content shared, namely the intent:
    - to normalise hateful or violent narratives,
    - to propagate hateful/violent conspiracies,
    - to foster hate or hostility towards an out-group.
  - Including implicit action triggers, such as the need for retaliation, need for self-defence, need to protect a group from reputational damage, instilling the idea of compensation/reward for undertaking action, exerting peer pressure, calling for respect for the code of honour

- **Tactic used to amplify the message or support wide distribution**
  - Reference names of prominent (political) figures.

5.1.2 Networking and Plotting

Ideological polarisation presents significant societal challenges, notably regarding its exploitation by terrorist and extremist groups to facilitate networking and coordination. The January 6 attack on the US Capitol is one of the most prominent cases of white supremacist and right-wing extremist online networking and plotting.[221] Already weeks prior to the attack, one could find plans of the attack circulating online, including on Facebook.[222] Beyond that, several extremist groups have demonstrated an ability to utilise the internet to propagate their ideologies, recruit members, and coordinate violent actions within their communities.

In general, anti-government extremism groups cultivate conspiracy theories through the sharing of common language, cultural heritage or geographical proximity in online groups.[223] These conspiracies are spread across many countries, and sometimes ideologies are adapted to the current context of the country. In the Netherlands, anti-government extremist groups fuelled by conspiracy narratives, particularly those influenced by the so-called 'sovereign citizen' ideology, have used online platforms to connect with like-minded networks across borders and seek international cooperation and coordination.[224] For example, individuals present on 'Common Law', the official website for the anti-government movement in the Netherlands, have been found to

219 Tanya Mehra, "Divided by Hate: Confronting Antisemitism and Islamophobia in the Netherlands," *Lawfare* (2025), https://www.lawfaremedia.org/article/divided-by-hate--confronting-antisemitism-and-islamophobia-in-the-netherlands.
220 See for instance posts #160, #161, #162, #164, and #166 of the OSINT research on file with the research team.
221 Williams et al., "The Online Extremist Ecosystem," 11.
222 Williams et al., "The Online Extremist Ecosystem," 11.
223 Bàrbara Molas et al., "Anti-Government Threats and their Transnational Connections," *ICCT* (March 2024): 29, https://www.icct.nl/sites/default/files/2024-06/FFO-Public%20Report%20English_Final%201.pdf.
224 Molas et al., "Anti-Government Threats and their Transnational Connections," 15.

have links to Telegram channels, which are used to coordinate and solicit people to participate in demonstrations and to display content from these demonstrations. The demonstration-focused movement on Telegram has previously been linked to large and disruptive protests in Canada and Australia, notably during the COVID-19 pandemic.[225] While calling for demonstrations in itself is fully legal (except when, for health care reasons, during the pandemic these gatherings were prohibited), calling for violence during demonstrations clearly is not.

Extremist groups are thus not solely plotting online but often also offline, as seen in the attempt to kill the Dutch Prime Minister or the failed coups d'état attempts prepared in both Germany and France.[226] Furthermore, a study identifies that the members involved in preparing these coups d'état attempts have a fascination with Russia and its president. Consequently, conspirators either directly contacted Russian representatives or pretended to have such connections to recruit others, who were all open-minded toward Russia.[227]

Among the posts collected for the OSINT research, there were many problematic references to historical or current context, which we consider to be a key indicator for implicit extremist content. Several posts reiterated Russian propaganda, one directly quoted Russian President Putin.[228] Some of the posts combine conventional anti-militarism and noninterventionist rhetoric with anti-institutional attitudes and false, misleading or ignorant information about the Russo-Ukrainian war and Russia's intentions behind the invasion of Ukraine,[229] whilst other posts connect Western support to Ukraine with antisemitic conspiracy theories.[230]

Another example of right-wing extremist online networking and plotting can be found in the Dutch Farmers' Movement. Far-right groups have utilised this movement to indoctrinate individuals against the government by using extremist discourse. Initially, the online discourse surrounding the farmers' movement focused on shared concerns about environmental policies.[231] However, the conversation online soon shifted to radical anti-government narratives, fuelled by disinformation, including claims about government intentions to steal farmland and other conspiracy theories.[232] This kind of rhetoric could play a role in the radicalisation towards anti-institutional extremism. The OSINT research also flagged several posts railing against climate policy and farmland reforms. They often contain conspiracy theories and false information regarding these policies.[233] Some posts push the narrative that the government is collaborating with climate activists to manufacture crises.[234]

In the Netherlands, fourteen individuals, many of whom are minors, were arrested in April 2025 on suspicion of inciting terrorism through social media platforms such as TikTok.[235] These arrests align with a broader pattern of youth-driven extremism in the Netherlands. In 2023 and 2024, several minors were arrested across the country for plotting attacks, many of them linked to far-right networks.[236] Since 2020, the Dutch right-wing extremist movement has largely been composed of minors, who now play a central role not only in the consumption but also in the production and the dissemination of extremist content, as well as in the plotting of attacks. Their

225 Molas et al., "Anti-Government Threats and their Transnational Connections," 30.
226 Molas et al., "Anti-Government Threats and their Transnational Connections," 32; AIVD, *Anti-institutional extremism in the Netherlands* (The Hague: Ministry of the Interior and Kingdom Relations, 2023), 7, https://english.aivd.nl/publications/publications/2023/11/7/publicatie-anti-institutional-extremism-in-the-netherlands.
227 Molas et al., "Anti-Government Threats and their Transnational Connections," 32.
228 See post #098 of the OSINT research, on file with the research team.
229 See posts #030, #061, and #202 of the OSINT research, on file with the research team.
230 See for instance post #036 of the OSINT research, on file with the research team.
231 Bàrbara Molas, "Dutch Flags and Maple Leaves: How Conspiracy Theories Created a Transnational Far-Right," *ICCT* (July 2024), https://icct.nl/publication/dutch-flags-and-maple-leaves-how-conspiracy-theories-created-transnational-far-right.
232 Molas, "Dutch Flags and Maple Leaves."
233 See posts #056, #082, #083, #084, and #089 of the OSINT Research, on file with the research team.
234 See posts #081, #086, and #095 of the OSINT research, on file with the research team.
235 "14 aanhoudingen in verband met opruiing tot terrorisme op sociale media," Openbaar Ministerie, https://www.om.nl/actueel/nieuws/2025/04/22/14-aanhoudingen-in-verband-met-opruiing-tot-terrorisme-op-sociale-media.
236 AIVD, *A web of hate: The online hold of extremism and terrorism on minors* (The Hague: Ministry of the Interior and Kingdom Relations, 2025), 12, https://english.aivd.nl/publications/publications/2025/04/03/a-web-of-hate.

prominent role within these circles is largely due to their digital skills and presence, and their content has been noticed by users both from the Netherlands and abroad.[237] The Dutch cases thus illustrate how young individuals are increasingly transitioning from consumers to active participants in the right-wing extremist environments.

Similarly, terrorist groups with other ideological backgrounds are using the online sphere for networking and plotting attacks. A study examined how a network of jihadist groups facilitates the dissemination of jihadi propaganda through online platforms.[238] The exploitation of social media by terrorist organisations enables the rapid and extensive distribution of jihadist content, ensuring its sustained presence in the digital sphere. It also allows these groups to tailor and target specific content to particular communities they aim to mobilise.[239] In 2024. for example, the Court of Appeal in The Hague sentenced a man to two years' imprisonment for participating in a terrorist organisation, incitement to terrorism, and training for terrorism after he was found guilty of disseminating violent, jihadist and ISIS material on social media channels, groups, and chats, including 30 Telegram channels promoting ISIS content.[240]

- **Indicator to identify implicit extremist content**
  - Presumed intent to cause harm by triggering to take offline (violent) action.

5.1.3 Recruitment and Mobilisation

Terrorist groups have increasingly leveraged the internet as a tool for recruitment and mobilisation, targeting individuals to advance their ideologies.[241] The growing engagement of youth with digital platforms has notably amplified their exposure to harmful and extremist content.[242] These groups exploit emerging technologies notably to radicalise young individuals who are susceptible to extremist propaganda and contribute to a cause perceived as larger than themselves.[243] Moreover, the internet facilitates the ability to reach a broader audience, enabling the identification and recruitment of individuals from vulnerable sectors of society. In particular, Hezbollah has been actively recruiting operatives from marginalised communities as well as organised criminal networks to further political objectives.[244] Additionally, these radicalised individuals have been mobilised to target Iran's adversaries globally, not only in Lebanon but also within the EU.[245]

Similarly, right-wing extremist groups made use of the internet and social media to recruit and mobilise members. In 2014, white supremacists launched a multi-year hate campaign known as *gamergate*.[246] Starting on Reddit, individuals were mobilised to target female gamers, female game developers, their male allies, men of colour, and feminists. Notably, this global campaign was not limited to online harassment but had real-life implications for the targeted individuals, including a famous American game developer, as their personal information was leaked online, and special forces were called to their homes.[247] This is once again an example of how online content can have offline consequences when a presumed intent to cause harm is included, for instance, by doxing.

237 AIVD, *A web of hate*, 3.
238 Ali Fisher, "Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence," *Perspectives on Terrorism* 9, no. 3 (2015): 3, https://pt.icct.nl/sites/default/files/import/pdf/2-swarmcast-how-jihadist-networks-maintain-a-persistent-online-presence-by-ali-fisher.pdf.
239 Fisher, "Swarmcast," 4.
240 Europol, *EU TE-SAT 2024*, 17.
241 Hunter et al., "The Metaverse as a Future Threat Landscape," 66.
242 Europol, *EU TE-SAT 2024*, 8.
243 Karen Greenberg, "Counter-Radicalization via the Internet," *The Annals of the American Academy of Political and Social Science* 668, no. 1 (2016): 1, https://doi.org/10.1177/0002716216672635.
244 Europol, *EU TE-SAT 2024*, 7.
245 Europol, *EU TE-SAT 2024*, 7.
246 Williams et al., "The Online Extremist Ecosystem," 7.
247 Williams et al., "The Online Extremist Ecosystem," 7.

Far-right activists in the UK have also long relied on an online-offline strategy for recruiting and mobilising people. The Patriotic Alternative (PA) in particular has developed an online communication that relies on more subtle messaging, mostly borderline content, to appeal to a broader audience and connect to offline activities as well.[248] Research has shown that the radicalisation process of individuals often involves a complex interaction between online and offline influences.[249] This suggests that individuals may be exposed to extremist content or ideologies through digital platforms, while simultaneously participating in real-world events or interactions that reinforce that messaging.[250]

Additionally, extremist recruiters often incorporate gaming elements in their propaganda to enhance user involvement, encouraging individuals to contribute by posting content or even livestreaming, which permits real-time interactions between creators and users.[251]

5.1.4 Financing

The rapid proliferation of social media has introduced a significant vulnerability in counter-terrorist financing regulations, as extremist and terrorist propaganda can be easily shared and used to encourage fundraising for specific causes.[252]

For example, researchers have highlighted that violent right-wing extremist networks are increasingly exploiting social media, crowdfunding platforms, and various online tools to facilitate the solicitation of donations and the financing of their activities.[253] The attack on the US Capitol on January 6, 2021, has drawn significant attention to the role of online crowdfunding platforms within extremist movements. Extremist groups have effectively utilised mainstream crowdfunding platforms to solicit financial support, thereby enabling them to reach large audiences and obtain substantial contributions.[254] These platforms, often facilitated by social media, act as conduits for connecting fundraisers with potential donors, amplifying the reach of their appeals. A particular focus of research has been the crowdfunding platform GiveSendGo, which has played a central role in financing the participants of the January 6 Capitol insurrection.[255] While the event was qualified as *terrorism* or *attempted coup*, individuals involved in the event, including those facing criminal charges, continue to receive financial support, raising over 5.3 million USD from more than 80,000 donors.[256]

Islamist extremist groups have also successfully combined religious messaging with social media tools to obtain financial support. This has enabled them to secure donations from both individuals intentionally supporting their cause and individuals who, through seemingly charitable platforms, unknowingly contribute to terrorist financing.[257] In Finland, authorities investigated a crowdfunding case involving nearly 400,000 EUR raised under the guise of charity and sent to terrorist groups in Syria.[258] Some funds were collected through crowdfunding platforms, however, many donors, primarily from Finland and abroad, were unaware of the potential terrorist connections.[259]

248 Allchorn, "Turning Back to Biologised Racism."
249 Allchorn, "Turning Back to Biologised Racism."
250 Chamin Herath, and Joe Whittaker, "Online Radicalisation: Moving Beyond a Simple Dichotomy," *Terrorism and Political* Violence 35, no. 5, (2021): 1039, https://doi.org/10.1080/09546553.2021.1998008.
251 Peeters et al., *Online extreemrechtse radicalisering* , 13-14.
252 Tom Keatinge, and Florence Keen, "Social Media and (Counter) Terrorist Finance: A Fund-Raising and Disruption Tool," *Studies in Conflict & Terrorism* 42, no. 1-2 (2019): 178, https://doi.org/10.1080/1057610X.2018.1513698.
253 Hans-Jakob Schindler, "Emerging challenges for combating the financing of terrorism in the European Union: financing of violent right-wing extremism and misuse of new technologies," *Global Affairs* 7, no. 5 (2021): 801, https://doi.org/10.1080/23340460.2021.1977161; Keatinge, and Keen, "Social Media and (Counter) Terrorist Finance," 192.
254 Rebecca Visser, "Crowdfunding Conspiracists: Grassroots Giving to January 6 Participants," *ICCT* (December 2024): 8, https://icct.nl/publication/crowdfunding-conspiracists-grassroots-giving-january-6-participants.
255 Visser, "Crowdfunding Conspiracists," 8.
256 Visser, "Crowdfunding Conspiracists," 1.
257 Keatinge, and Keen, "Social Media and (Counter) Terrorist Finance," 199.
258 FATF, *Crowdfunding for Terrorism Financing* (Paris: FATF, 2023), 34, https://www.fatf-gafi.org/content/dam/fatf-gafi/reports/Crowdfunding-Terrorism-Financing.pdf.coredownload.inline.pdf.
259 FATF, *Crowdfunding for Terrorism Financing*, 34.

Additionally, the misuse of cryptocurrency by extremist groups has emerged as an evolving threat.[260] Due to their decentralised and anonymous nature, transactions with cryptocurrencies have become a preferred method for financing terrorism. Research indicates that the variety of virtual currencies complicates regulation, as existing legal definitions focus mainly on blockchain-based currencies and exclude those issued by central banks or governments.[261]

5.2 Use of Advanced Technologies

Many terrorist and extremist groups have adopted new tactics that leverage cutting-edge digital technologies, such as large language models (LLMs), artificial intelligence (AI), deepfakes, and video games, to evade detection.[262] Additionally, research has found that extremist actors online deploy different tactics, such as the use of humour, coded language, and misleading images to evade detection and spread their narratives to a wide audience.[263] Such tactics are often an indication of users' intent to hide the true meaning of the content.

The increasing development of technological tools has benefited extreme right groups by enabling them to easily spread their ideologies online. Looking at right-wing populist fake social media accounts and image-based means of online communication in Germany, researchers found that this image-based content is often emotionally appealing to the audience, and eventually has the potential to influence democratic processes. They concluded that image based-content is more likely to be shared online, easier to consume for the audience, and more emotionally engaging than text-based content.[264]

The OSINT research conducted for this study also yielded some examples of image-based content designed to trigger emotional engagement. For instance, one post simplifies the idea of so-called race-mixing by using the analogy of mixing colours in a washing machine.[265]

Furthermore, researchers found that Facebook's algorithm favours image-based content, giving it a wider reach, which is particularly problematic when such content is shared by fake accounts.[266] As seen from fake social media profiles affiliated with the German right-wing party Alternative für Deutschland (AfD), fake accounts can successfully bypass automated and human based detection for a significant time. By varying the timing and content of posts, these accounts avoid automated spam-detection systems. They combine different images, including misleading cover images or blurred or otherwise altered images, with slightly altered or misspelt text, and the ambiguous or coded language to evade pattern recognition algorithms. Additionally, they often mimic genuine user behaviour by engaging sporadically in real-time interactions. This mix of automated and human-like actions makes it harder for algorithms and real users to distinguish fake profiles from real ones, thus prolonging their activity on the platform.[267]

Additionally, AI and LLM tools have been increasingly used by terrorist and extremist groups to spread their ideologies. This adaptation to advanced technologies has led to the rise of new forms of information warfare, propaganda systems, and disinformation campaigns.[268] One particular concern is the potential misuse of LLMs by extremists for propaganda purposes, as

260 Schindler, "Emerging challenges for combating the financing of terrorism in the European Union," 803.
261 Annelieke Mooij, *Regulating the Metaverse Economy: How to Prevent Money Laundering and the Financing of Terrorism* (Cham: Springer, 2024), 109-112, https://doi.org/10.1007/978-3-031-46417-1.
262 Europol, *EU TE-SAT 2024*, 6.
263 Molas, "Alt-solutism."; Fielitz, and Ahmed, "It's not funny anymore."
264 Stephen Albrecht, and Merle Strunk, „Memes für die Massen: Rechtspopulistische Fake-Accounts und ihre visuellen Strategien," in *Bilder, soziale Medien und das Politische: Transdisziplinäre Perspektiven auf visuelle Diskursprozesse* (Bielefeld: transcript Verlag, 2021), 149-180, https://doi.org/10.1515/9783839450406-007.
265 Post #003 of the OSINT research, on file with the research team.
266 Albrecht, and Strunk, „Memes für die Massen," 154-155.
267 Albrecht, and Strunk, „Memes für die Massen," 156-159.
268 Stephane Baele et al., "Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey," *Terrorism and Political Violence* (2024): 1, https://doi.org/10.1080/09546553.2024.2380089.

these tools facilitate the rapid production of extremist propaganda by fewer individuals.[269] In particular, ISKP has been capitalising on AI and chatbots to create and spread propaganda in different languages.[270] Despite efforts by big tech companies to regulate these advancements, the technology is becoming more accessible and less controlled, as illustrated by the 2016 incident involving Microsoft's AI model, Tay, which posted extreme-right content.[271] An ever more concerning development is how chatbots can encourage often lonely young people to commit a terrorist attack. JS Chail attempted the assassination of Queen Elizabeth II after exchanging over 5000 messages with a chatbot named Sarai.[272]

Deepfakes have also increasingly been employed as a tool to incite political violence. Deepfake videos are strategically utilised to foster societal unrest, erode public trust in democratic institutions, and amplify politically polarising narratives.[273] This emerging phenomenon poses a significant risk of fuelling disinformation and perpetuating the spread of conspiracy theories. Some scholars project that by 2026, up to 90 percent of online content may be synthetically generated, suggesting that deepfakes could become a widespread source of cybercrime. [274]

A key concern for researchers is the public's general unawareness of the potential dangers posed by deepfakes.[275] Due to their hyper-realistic nature, these videos are exceedingly difficult to distinguish from legitimate media, which poses a challenge not only to human discernment but also to automated detection systems.[276] Additionally, certain groups, perceiving themselves as defenders against perceived governmental overreach, may exploit deepfake technology to create content portraying the government as corrupt or tyrannical, further complicating efforts to maintain the integrity of public discourse.[277] In addition to deepfakes, individuals affiliated with far-right groups are leveraging AI to disseminate harmful content, facilitate networking among like-minded individuals, and solicit funding.[278] These actors frequently engage in collaborative efforts within groups, exchanging strategies to bypass safety protocols and generate far-right disinformation. This activity contributes to the widespread normalisation of radicalising material on widely accessible platforms.

Furthermore, gaming platforms possess the capacity to engage users by simulating real-world scenarios within virtual environments, thereby amplifying the influence of such messages. Hence, video games have emerged as a new tool for disseminating jihadi propaganda.[279] The universal nature of video games provides significant opportunities for the transmission of jihadist ideologies, easily reaching billions of individuals.[280] The gaming platform Minecraft with over 200 million monthly active players, is also being exploited by right-wing groups as a tool for radicalisation.[281]

Once individuals have been recruited, they are then encouraged to subscribe to alternative social media platforms, in case of increased content moderation on more mainstream services.[282]

269 Stephane Baele et al., "Is AI-Generated Extremism Credible?" 3.
270 Mona Thakkar, and Anne Speckhard, "Caliphate AI - IS/ISKP Supporters Harness GenerativeAI for Propaganda Dissemination," *International Center for the Study of Violent Extremism* (July 2024), https://www.researchgate.net/publication/382025204_Caliphate_AI_-ISISKP_Supporters_Harness_Generative_AI_for_Propaganda_Dissemination.
271 Stephane Baele et al., "Is AI-Generated Extremism Credible?" 4.
272 Priyank Mathur et al., "The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence," *ICCT* (May 2024), https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence.
273 Ella Busch, and Jacob Ware, "The Weaponisation of Deepfakes: Digital Deception by the Far-Right," *ICCT* (December 2023): 2, https://www.icct.nl/sites/default/files/2023-12/The%20Weaponisation%20of%20Deepfakes.pdf.
274 Busch, and Ware, "The Weaponisation of Deepfakes," 3.
275 Busch, and Ware, "The Weaponisation of Deepfakes," 3.
276 Busch, and Ware, "The Weaponisation of Deepfakes," 8.
277 Busch, and Ware, "The Weaponisation of Deepfakes," 5.
278 Busch, and Ware, "The Weaponisation of Deepfakes," 7; Williams et al., "The Online Extremist Ecosystem," 9.
279 Miron Lakomy, "Let's Play a Video Game: Jihadi Propaganda in the World of Electronic Entertainment," *Studies in Conflict & Terrorism* 42, no. 4 (2019): 383, https://doi-org.ezproxy.leidenuniv.nl/10.1080/1057610X.2017.1385903.
280 Lakomy, "Let's Play a Video Game," 385.
281 Gagandeep, "Playing with Hate: How Far-Right Extremists Use Minecraft to Gamify Radicalisation," *GNET Insights* (2025), https://gnet-research.org/2025/07/02/playing-with-hate-how-far-right-extremists-use-minecraft-to-gamify-radicalisation/.
282 Allchorn, "Turning Back to Biologised Racism."

Alternative social media platforms benefit from minimal or absent content moderation, which allows greater opportunities for the online proliferation of hate speech and violent extremist rhetoric.[283] Telegram, in particular, has emerged as a prominent platform for extremist groups.[284] These platforms serve as tools for recruitment and for fostering strong communities where participants can actively engage.

Several posts collected during the OSINT research illustrate this tactic. For instance, several posts by the same user refer to a Discord server called 'Fighters for Freedom'.[285] Another user posts links to his own publications that are often far more inflammatory than the content they post on Instagram.[286] Several users further posted links to niche and far-right publications.[287]

Finally, research has examined the similarities and disparities in the ways in which right-wing extremists and jihadi extremists utilise the internet.[288] While both groups leverage new technologies, the content they produce diverges significantly. Violent jihadi extremists tend to employ language that emphasises social and emotional processes, along with religious themes, more frequently than their right-wing extremists. In contrast, right-wing extremists are more inclined to employ language that evokes anxiety, while refraining from promising an idealised or prosperous future.[289]

- **Indicators of implicit extremist content**
  - Methods to conceal meaning by:
    - using a misleading cover image
    - using blurred/altered images

- **Tactics/mechanisms used to amplify the message or support wide distribution**
  - Use of new technology, such as LLMs, AI, bots, deepfakes, and videogames.
  - Tactics to bypass moderation/detection, such as the use of fake accounts, altered text, or mimicking genuine user behaviour
  - Tactics for further amplification/diversification of platforms
    - Hiding behind anonymity to lower the thresholds to post content
    - Referring to other platforms within the post, or by copying the post to other platforms.

5.3 Use of Attractive Formats

Research further found that while the content of the image-based communication is provocative and explicit, they try not to offend the general audience and rather mainstream their hateful narratives. Hence, this content is less extreme and grave than similar content shared within right-wing bubbles to ensure that it is not flagged by non-likeminded users. These memes are often referred to as "memes for the masses".[290] Yet, it is well understood by likeminded users, supporting their world view and ideologies.[291]

Examples of "memes for the masses" were also found among our own OSINT research. One post representing a collage of images suggesting a Jewish conspiracy behind key moments in history, with a text stating "never forget what they have done", a David star, and hashtag referencing

283 Samantha Walther, and Andrew McCoy, "US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism," *Perspectives on Terrorism* 15, no. 2 (April 2021): 101-104, https://www.icct.nl/sites/default/files/import/publication/walther-and-mccoy-.pdf.
284 Stoeldraaijers et al., *Radicale reclame op sociale media*, 6.
285 Posts #003, and #005-008 of the OSINT research, on file with the research team.
286 Posts #174, #180, and #183 of the OSINT research, on file with the research team.
287 Posts #031, #038, #072, #097, #100, and #233 of the OSINT research, on file with the research team.
288 Weeda Mehran et al., "Two Sides of the Same Coin? A Largescale Comparative Analysis of Extreme Right and Jihadi Online Text(s)," *Studies in Conflict & Terrorism* (2022): 18, https://doi.org/10.1080/1057610X.2022.2071712.
289 Mehran et al., "Two Sides of the Same Coin?" 18.
290 Albrecht, and Strunk, „Memes für die Massen," 174-175.
291 Albrecht, and Strunk, „Memes für die Massen," 174-175.

with "third position" to the Third Reich, and with "88", representing two times the 8th letter of the alphabet.[292] Another example was found with a meme of a green star of the Moroccan flag chasing the blue David star.[293]

Another tactic applied by extremist actors online, in particular from the far-right spectrum, to evade content detection while at the same time appealing to a broader audience is the use of humour.[294] Similar observations have been made in the Netherlands, where Dutch far-right groups use various online platforms to disseminate memes that promote extremist narratives. According to the NCTV, more veiled and 'humorous' memes are considered to be more convenient than direct calls for hatred and violence as they can bypass content detection mechanisms. They have proved easy to disseminate on mainstream online platforms, where particularly interested users are then encouraged to join alternative or private platforms.[295]

Similarly, humorous memes are increasingly used by extreme right groups to spread their ideas in an attempt to reach a broader audience and recruit new people.[296] For instance, in 2017, the militant neo-Nazi organisation, the Nordic Resistance Movement (NRM), was sending weekly online content relevant to the organisation, which included a "Meme of the week".[297] This incorporation of humour and irony shows the shift in extremist groups' online communication.
OSINT research conducted for this study also detected several harmful humorous posts, ranging from problematic 'memes for the masses' to more inflammatory memes that often involve extremist narratives or conspiracies targeting minorities. There are posts ridiculing COVID-measures,[298] distrustful of or conspiratorial about democratic governments and independent media,[299] posts against multiculturalism and race-mixing or non-white people more generally,[300] and posts affirming the International Jewish conspiracy theory and other related antisemitic conspiracy theories.[301]

This new use of ambiguous content pushes the boundaries between hate speech and freedom of expression, as will further be assessed below. Memes, which are easily shared and remixed via digital media, are now being used to spread far-right views.[302] They exist in a grey area, where it is hard to tell whether they are meant to be taken seriously or are simply a form of mockery. This confusion complicates the relationship between extremism, satire, and intent.

- **Indicators of implicit extremist content**
  - Methods to conceal meaning by:
    - using ambiguous/coded language
    - using altered text or misspelling of word
    - using humour or irony

- **Tactics/mechanisms used to amplify the message or support wide distribution**
  - Tailored content to trigger click baits and further sharing, such as image based content, memes, memes for the masses.

292 Post #006 of the OSINT Research, on file with the research team.
293 Post #120 of the OSINT Research, on file with the research team.
294 Fielitz, and Ahmed, "It's not funny anymore," 5-6.
295 NCTV, *Memes als online wapen: Fenomeenanalyse naar het gebruik van memes door extreemrechts* (The Hague: Ministry of Justice and Security, 2024), 11-12, https://www.nctv.nl/documenten/publicaties/2024/05/21/fenomeenanalyse-memes-als-online-wapen.
296 Fielitz, and Ahmed, "It's not funny anymore," 6.
297 Tina Askanius, "On Frogs, Monkeys, and Execution Memes: Exploring the Humor-Hate Nexus at the Intersection of Neo-Nazi and Alt-Right Movements in Sweden," *Television & New Media* 22, no. 2 (February 2021), https://doi.org/10.1177/1527476420982234.
298 Post #060 of the OSINT research, on file with the research team.
299 Posts #036, #092, and #107 of the OSINT research, on file with the research team.
300 Posts #003, and #024 of the OSINT research, on file with the research team.
301 Posts #002, and #019 of the OSINT research, on file with the research team.
302 Molas, "Alt-solutism."

5.4 Gaps and Observations

This chapter has shown that while research examining the similarities and differences in how right-wing extremists and violent jihadi extremists use the internet was initially lacking, it has now expanded.[303] However, such analyses frequently maintain a narrow focus on either jihadism or right-wing extremism without a comparative lens.

Since initial research has established a connection between offline events such as the Israel/Gaza conflict or the war in Ukraine, and the ability of terrorists and extremists to exploit these events for their purposes in the online sphere,[304] further analysis of the impact of these events on the radicalisation process, including the study of different countries and ideologies, is needed. Although there have been regulatory steps at the EU level to combat online terrorism financing, such as controlling online content and addressing cryptocurrency misuse, the regulation of more complex technologies, including privacy coins, non-custodial wallets, and mixers, appears inadequate.[305] Notably, gaps have been identified in the regulation of cryptocurrency layering within the Metaverse economy.[306] The Financial Action Task Force Report on Crowdfunding for Terrorism Financing has further highlighted the challenges faced by states in addressing these emerging threats. The report underscores a current lack of expertise and understanding of terrorist financing and violent extremism through online platforms.[307] Finally, research has identified a significant gap in the documentation and rigorous analysis of terrorist financing via social media, which presents a critical gap and vulnerability.[308]

Ultimately, the lack of harmonised terminology such as radicalisation, extremism, terrorism and borderline content, along with the ambiguous scope of these notions, creates significant challenges in analysing the current literature and making meaningful comparisons between studies.[309]

303 Mehran et al., "Two Sides of the Same Coin?" 18.
304 Völker, "How terrorist attacks distort public debates," 3509; Pauwels, and Van Alstein, "Polarisation," 10-11.
305 Schindler, "Emerging challenges for combating the financing of terrorism in the European Union," 806.
306 Mooij, *Regulating the Metaverse Economy*, 112.
307 FATF, *Crowdfunding for Terrorism Financing*, 31-33.
308 Keatinge, and Keen, "Social Media and (Counter) Terrorist Finance," 199.
309 Van Wonderen et al., *Rechtsextremisme op sociale mediaplatforms?*

6. Policies and Practices of Detection and Moderation

This chapter will examine the various policies that online service providers use to detect and moderate harmful content. The primary focus is on how reliable these detection systems are and whether unreliable methods lead to unnecessary content removal decisions and subsequent restrictions of human rights.

This chapter will first elaborate on the existing detection and moderation policies and practices, which are often interconnected. It will examine the use of trusted flaggers, hashtag databases, and oversight mechanisms. Following this overview, the analysis turns to the reliability of detection and content moderation. When moderating online content, hosting service providers (HSPs) must adequately protect freedom of expression and adopt rule-of-law-compliant policies and practices in full respect of human rights.[310] The relevant section in this chapter will explore how unreliable detection can violate the right to free speech through false positives. Conversely, it will also examine false negatives, in particular regarding implicit extremist content.

6.1 Detection Policies and Practices

Recent data suggests that, in light of strict regulatory frameworks that require platforms to delete illegal, terrorist, or other harmful content (such as hate speech, CSAM, or fraud) within short periods and imposing fines for failure to do so, many HSPs deploy automated detection and moderation tools.[311] Especially bigger HSPs, such as Meta and Google, have started using AI-based tools in detecting and moderating content that violates their own Community Guidelines (CG). A study on content moderation of HSPs falling under the DSA confirms this observation, as it shows that most of the content moderation decisions that were reported were made by very large online platforms (VLOPs), with Google Shopping being responsible for more than half of the submitted Statements of Reasons (SoR).[312] The data further indicates that while individual platforms such as X, Booking.com, and LinkedIn do not make any automated decisions on content moderation, several platforms, including TikTok, rely on fully automated decision making in content moderation.[313] The vast majority (99.8 percent) of all content moderation reported in the DSA Transparency Database as of November 2023, relates to violations of the individual service's Terms of Service (ToS), rather than the illegality of the content, which was only cited in 0.2 percent of all reported content moderations.[314] This highlights the crucial role that HSPs play in content moderation as well as the shattered landscape of definitions and regulatory frameworks that differ between HSPs.

Many HSPs deploy a combination of automated tools and human scrutiny. For example, in cases in which the classification of content is unclear, human reviewers assess the individual content to decide whether it has to be subject to moderation or not.[315] Additionally, the detection of content can be conducted by automated means with human scrutiny at the decision making level, meaning the content has already been classified, and humans merely decide on which means of moderation should be taken.[316] Larger HSPs, such as Facebook or YouTube, for example, employ

310 Stuart Macdonald et al., "Regulating terrorist content on social media: automation and the rule of law," *International Journal of Law in Context* 15, no. 2 (2019): 185 - 186, https://doi.org/10.1017/S1744552319000119.
311 Natalie Alkiviadou, "Platform liability, hate speech and the fundamental right to free speech," *Information & Communications Technology Law* 34, no. 2 (2024): 2, https://doi.org/10.1080/13600834.2024.2411799.
312 Kaushal et al., "Automated Transparency," 1124
313 Kaushal et al., "Automated Transparency," 1124. One must note that the decision-making process is separate from the preceding detection of to be moderated content, which is done through (partially) automated means.
314 Kaushal et al., "Automated Transparency," 1124-1125.
315 Federico Galli et al., "The Regulation of Content Moderation," in *The Legal Challenges of the Fourth Industrial Revolution* (Cham: Springer, 2023), 67, https://doi.org/10.1007/978-3-031-40516-7_5.
316 Kaushal et al., "Automated Transparency," 1124-1125.

dedicated teams of reviewers assessing between 600 and 800 pieces of content each day.[317] Instagram takes a similar approach, stating that content is generally detected by AI tools and only sent for human review to decide on adequate moderation means when it is not already clearly identified as a violation of the CG but still identified as potentially "inappropriate, disrespectful or offensive."[318] Similarly, Reddit uses an AI tool for content detection and subsequent automatic detection in case of clear breaches of the Reddit Rules.[319]

While detection and eventual moderation of content are separate processes, studies have found a strong correlation between fully automated detection and fully automated decision making, meaning providers that employ fully automated tools to detect content usually also employ fully automated tools to decide on the applicable means of content moderation for the detected content.[320]

However, automatised detection mechanisms are criticised for leading to false positives, meaning flagging and potential subsequent deletion of lawful content, since they often fail to recognise linguistic specifics and individual context.[321] A 2024 study found that between 87.5 percent and 99.7 percent of the content removed on Facebook or YouTube in France, Germany, and Sweden was legally permissible. [322]

Additionally, automated tools are usually trained on the most common languages, hence failing to efficiently detect harmful content in less common languages.[323] Furthermore, small or micro sized HSPs with limited financial and human resources often lack the capacity to adequately detect and moderate content by manual means or to develop and deploy automated tools, as also confirmed by experts interviewed for this study.[324] Complementing the emergence of companies dedicated to the development of content detection and moderation tools, in particular in relation to terrorist content, the EU funds several projects aimed at supporting small and micro-sized HSPs to conduct automated content moderation to fulfil their obligations under the TCO.[325]

Studies examining the automated dissemination of harmful content online found that flagging – meaning users reporting critical content to the platforms – presents a key aspect of effective content moderation.[326] This is particularly relevant for certain types of platforms, such as gaming platforms, where due to the interactive and often hidden nature of user communication other content detection and moderation means are insufficient to address the full scale of harmful content.[327] However, flagging systems of many major platforms need to be improved, as only ten percent of the content that they themselves have flagged on the platforms was eventually removed or disabled after eight weeks.[328] Additionally, a recent analysis of platforms' content flagging and reporting mechanisms shows that platforms, in particular TikTok, alongside other

317 Philipp Schneider, and Marian-Andrei Rizoiu, "The effectiveness of moderating harmful online content," *PNAS* 120, no. 34 (2023): 2, https://www.pnas.org/doi/10.1073/pnas.2307360120.
318 "How Instagram uses artificial intelligence to moderate content," Instagram, accessed August 12, 2025, https://help.instagram.com/423837189385631/?helpref=related_articles.
319 "Content Moderation, Enforcement, and Appeals," Reddit, accessed August 12, 2025, https://support.reddithelp.com/hc/en-us/articles/23511059871252-Content-Moderation-Enforcement-and-Appeals.
320 Kaushal et al., "Automated Transparency," 1124-1125.
321 Alkiviadou, "Platform liability, hate speech and the fundamental right to free speech," 7-8.
322 The Future of Free Speech, *Preventing "Torrents of Hate" or Stifling Free Expression Online?* 6.
323 Erich Prem, and Brigitte Krenn, "On Algorithmic Content Moderation," in *Introduction to Digital Humanism* (Cham: Springer, 2023), 488, https://doi.org/10.1007/978-3-031-45304-5_30.
324 George Kalpakis et al., "AI-Based Framework for Supporting Micro and Small Hosting Service Providers on the Report and Removal of Online Terrorist Content," in *Paradigms on Technology Development for Security Practitioners* (Cham: Springer, 2025), 250-251, https://link.springer.com/book/10.1007/978-3-031-62083-6; interview conducted on 11 June 2025, on file with research team.
325 Kalpakis et al., "AI-Based Framework for Supporting Micro and Small Hosting Service Providers on the Report and Removal of Online Terrorist Content," 249-261.
326 Carpani et al., *EU Internet Forum*, 51-55.
327 Galen Lamphere-Englund, and Menso Hartgers, *CTRL+ALT+COLLABORATE: Public-Private Partnerships to Prevent Extremism in Gaming* (Luxembourg: Publications Office of the EU, 2024), 36-40, https://home-affairs.ec.europa.eu/document/download/e446f013-34e1-4f74-bce6-90f661937ce9_en; Menso Hartgers, and Eviane Leidig, "Fighting extremism in gaming platforms: a set of design principles to develop comprehensive P/CVE strategies," *ICCT* (June 2023), https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it; Albrecht, and Strunk, "Memes für die Massen," 176.
328 Carpani et al., *EU Internet Forum*, 43.

VLOPs, design their flagging and reporting mechanisms not always in compliance with the DSA.[329] Instead of following the procedures in Article 16 DSA on how to handle flagged content in a timely and appropriate manner, their reporting mechanisms merely allow users to report content based on the platform's own terms of service, instead of referring to allegedly illegal content as outlined in the DSA.[330]

Triple verification: The Dutch Authority for the prevention of online Terrorist Content and Child Sexual Abuse Material (ATKM) conducted a study to explore whether triple verification of CSAM and terrorist content would be useful and result in more accurate and reliable assessments. The use of hash (sharing) databases can help platforms to quickly identify whether certain content is already identified as terrorist, CSAM, or otherwise harmful content and take appropriate measures. However, incorrect matches could lead to false positives, which would infringe on the freedom of expression or to false negatives, in which case harmful content remains online. The study revealed that the number of verifications should take the complexity of the content into account. Content that is clearly legal or illegal could be assessed by two assessments.[331]

6.2 Moderation Policies and Practices

Challenges faced by HSPs in relation to content moderation are rapidly evolving, necessitating continuous updating of policies and practices. Content moderation policies and practices must be tailored to the specific legal, cultural, and social contexts in which content is published, ensuring that they reflect the diverse nature of online platforms.

There are various ways in which VLOPs and HSPs can share their policies on content moderation, ranging from ToS and CG to content reporting guidelines.[332] Before the implementation of the TCO and DSA, research into content moderation policies concerning harmful content showed that most HSPs did not provide sufficient information on applicable definitions, identification, reporting, and decision making procedures, and available remedies.[333] Recent research indicates that this did not seem to have changed since the entering into force of the TCO and the DSA, since only 25 percent of the HSPs in France, Germany, and Sweden provide public information on their content moderation policies.[334] This makes it difficult for users to assess what exact policies apply.[335] At the same time, many providers use internal policies and guidelines in conducting content moderation.[336] This lack of transparency and knowledge gap is particularly concerning since the predictability of content moderation not only allows users to easily identify potential breaches of the policies and appeal moderation decisions, but also to self-regulate their behaviour online.[337]

As noted by the Council of Europe Steering Committee for Media and Information Society (CDMSI) in 2021, content moderation policies require constant review and adaptation to respond

329 "Follow Me to Unregulated Waters!" Holznagel.
330 Holznagel, Daniel, "Follow Me to Unregulated Waters!: Are Major Online Platforms Violating the DSA's Rules on Notice and Action?," *VerfassungsBlog*, May 30, 2024, https://doi.org/10.59704/80267b8bd7a278a4.
331 Melissa Rottier, *Onderzoek naar de verificatie van kinderpornografisch en terroristisch materiaal ten behoeve van databases* (Rotterdam: ATKM, 2025), https://www.atkm.nl/documenten/2025/07/23/onderzoek-naar-kwaliteit-databases-die-moeten-voorkomen-dat-schadelijk-materiaal-wordt-gedeeld.
332 Sabine Einwiller, and Sora Kim, "How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation," *Policy & Internet* 12, no. 2 (2020): 193, https://doi.org/10.1002/poi3.239.
333 Einwiller, and Kim, "How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication," 202.
334 The Future of Free Speech, *Preventing "Torrents of Hate" or Stifling Free Expression Online?* 7.
335 The Future of Free Speech, *Preventing "Torrents of Hate" or Stifling Free Expression Online?* 7.
336 Mchangama et al., *Scope Creep*, 16-17.
337 CDMSI, *CONTENT MODERATION: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation* (Strasbourg: Council of the EU, 2021), 6-7, https://edoc.coe.int/en/internet/10198-content-moderation-guidance-note.html.

to technological evolutions and emerging trends in the dissemination of content online.[338] Regardless, HSPs should make proportionate decisions in moderating content and not generally delete content or even deplatform users, but rather consider the entire range of available means.[339] However, Reddit lays out that in moderating content that violates the Reddit Rules, they apply an upscaling approach. This means that when content is identified as a breach of the Reddit rules, it will not automatically be removed. Rather, the moderation decision is made on a case-by-case basis, taking into account, among others the violation history of the user, the gravity, and the type of content.[340] While TikTok is also employing teams for human review of automatically detected content,[341] announcements in mid-2025 to lay off a significant number of human reviewers for the German speaking market, covering more than 30 million users, after already dissolving the team of around 300 Dutch speaking human reviewers in September 2024, indicate a shift away from human review in content moderation.[342]

The Trust and Safety Professionals Association (TSPA) provides an overview of different means of moderation available to hosting service providers, depending on the type of abuse.[343] Some forms of abuse, however, require immediate action. This could be the case with CSAM or if there is an imminent threat to life.

Table 2: Overview of different content moderation measures originally prepared by Eric Goldman, co-founder of TSPA and Professor of Law at Santa Clara University[344]

Content Regulation	Account Regulation	Visibility Reductions	Monetary	Other
Remove content	Terminate account	Shadow ban	Forfeit accrued earnings	Educate users
Suspend content	Suspend account	Remove from external search index	Terminate future earnings (by item or account)	Assign strikes/ warnings
Relocate content	Suspend posting rights	No-follow author's links	Suspend future earnings (by item or account)	Outing/unmasking
Edit/redact content	Remove credibility badges	Remove from internal search index	Fine author/impose liquidated damages	Report to law enforcement
Interstitial warning	Reduce service levels (data, speed, etc.)	Downgrade internal search visibility		Put user/content on blocklist
Add warning legend	Shaming	No auto-suggest		Community service
Add counter speech		No/reduced internal promotion		Restorative justice/ apology
Disable comments		No/reduced navigation links		
		Reduced virality		

338 CDMSI, *CONTENT MODERATION*, 4.
339 CDMSI, *CONTENT MODERATION*, 5-6.
340 "Content Moderation, Enforcement, and Appeals," Reddit.
341 "Content Moderation," TikTok, accessed August 12, 2025, https://www.tiktok.com/euonlinesafety/en/content-moderation/.
342 Stijn Bronzwaer, "Complete moderatieteam TikTok in Nederland ontslagen," *NRC*, published October 14, 2024, https://www.nrc.nl/nieuws/2024/10/14/complete-moderatieteam-tiktok-in-nederland-ontslagen-a4869230; Dara Kerr, "TikTok to replace trust and safety team in Germany with AI and outsourced labor," *The Guardian*, published August 10, 2025, https://www.theguardian.com/technology/2025/aug/10/tiktok-trust-safety-team-moderators-ai.
343 "Enforcement Methods and Actions," TSPA, accessed January 16, 2025, https://www.tspa.org/curriculum/ts-fundamentals/policy/enforcement-methods/.
344 "Enforcement Methods and Actions," TSPA.

		Age-gate		
		Display only to logged-in readers		

Each of these moderation measures – when lawfully applied – restricts the freedom of expression. The severity of these measures regarding impact on the freedom of expression is thereby declining from the most severe measures listed at the top to the less severe measures listed at the bottom. For example, shaming may impose less of a restriction than permanently deleting the account of a user.

At the same time, it is important to distinguish between measures that are directed at the content and those that are directed at individual users. While it may at times be difficult to identify the type of content, attributing the creation, possession or dissemination to an individual and taking subsequent steps is even more difficult. The problem is that the *mens rea* element of a criminal act (intent) differs across offences. Intent may sometimes be explicitly included in the criminal offence or may be inferred. Under Dutch criminal law, different forms of intent ("opzet") and culpability ("schuld") exist, which determine the extent to which an individual deliberately acted or should have acted. These different forms of intent and culpability are important in assessing how a criminal offence can be attributed to an individual. While platforms do not need to establish that a criminal offence has been committed like a court, the different types and gradations of the mental element should nevertheless guide the coder which moderation decision should be taken against an individual.

Unclear definitions of what constitutes illegal content, particularly in relation to borderline content and so-called hate speech, also pose challenges for HSPs in developing nuanced content moderation systems that allow them to adhere to the obligations under the DSA while at the same time not over-blocking content.[345] As the DSA is relatively new, it remains to be seen whether HSPs will find nuanced approaches or err on the side of caution towards DSA obligations by over-blocking. The transparency obligations outlined below, and in particular data access for researchers, could be useful tools to establish scrutiny over HSPs content moderation practices in this regard. The CDMSI also notes that each category of online content, for example, terrorist, illegal, implicit extremist content, should be addressed by separate policies that are targeted towards the precise category and type of content. Nevertheless, some platforms, such as Reddit, state in their moderation policies that they apply a fully automated detection and moderation approach to "terrorist, CSAM, and other non-consensual intimate imagery." If the automated detection finds a high probability that such content violates the Reddit Rules, it is not being forwarded for human review but immediately removed.[346]

The overall lack of clarity on and inaccessibility of content moderation policies is in stark contrast to the fact that many providers have indeed implemented dedicated measures to address terrorist and other harmful content on their services, including the use of automated detection and moderation tools as described above.[347] Instagram, for example, deploys an AI tool to detect and remove all content that violates the Community Guidelines. In cases of unclear categorisation, content on Instagram is being forwarded to human reviewers. However, their categorisation of the content and subsequent decision on an adequate means of moderation is then fed into the machine learning system, which further develops the AI detection and moderation tool.[348] Subsequently, human review could become less if the AI tool increasingly detects and moderates content based on human feedback from previously AI detected content.

345 Therese Enarsson, "Navigating hate speech and content moderation under the DSA: insights from ECtHR case law," *Information & Communications Technology Law* 33, no. 3 (2024), 400-401, https://doi.org/10.1080/13600834.2024.2395579.
346 "Content Moderation, Enforcement, and Appeals," Reddit.
347 European Commission, *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL*, 8.
348 "Content Moderation, Enforcement, and Appeals," Reddit.

Furthermore, Reddit's user-based approach to content moderation, with little independent overview on removal decisions can even foster ideological echo chambers. Research into deleted content on Reddit has shown that user moderators are more likely to moderate content in accordance with their own political views, thus creating spaces aligned with these views.[349] This can become particularly problematic in relation to extremist and implicit extremist content for the reasons described above in chapter 5.

6.3 Reflecting on Practices of Online Service Providers

Different platforms have different approaches to content moderation. This relates to the different types of services they offer and their general willingness to engage in content moderation. Concerning the latter, a 2023 study found that terrorists rely on a wide range of different types and sizes of online platforms to spread their content.[350] These include social media platforms, streaming services, file sharing and storage services, as well as video, image, and audio sharing services. Further, terrorists also rely on messenger services, book subscriptions, and, to a lesser extent, search engines and link shorteners.[351] Notably, some messenger platforms on which the researchers of the 2023 study found terrorist content are run by terrorist groups themselves or their supporters.[352] In general, it can be found that some new social media platforms and messenger services advertise little to no content moderation on their platforms, attempting to attract users who do not want their content to be scrutinised.[353]

As for the varying moderation approaches across different types of platforms, recent research found that not all platforms moderate the same type of content (e.g. text, video, images) to the same extent but rather focus on the type that is most prominent on their service. For example, TikTok predominantly moderates text and videos, while Snapchat predominantly moderates videos and images.[354] Those platforms that do moderate content seem to predominantly rely on content removal or disabling access thereto.[355] However, when it comes to so-called borderline content – identified by the platforms as content that is close to violating their ToS platforms most commonly resort to reducing public visibility of the content, according to studies.[356]

At the same time, researchers point out that domestic legal proceedings on alleged violations of freedom of speech through the removal of alleged hate speech take several years, while hosting service providers are facing strict deadlines to remove alleged terrorist or harmful content, such as the one-hour deadline under the TCO and the undue delay requirement under the DSA.[357] While this suggests an unfair burden on hosting service providers and the impossibility of proper legal assessments by them, it fails to acknowledge that these tight deadlines merely apply in relation to take-down requests issued by competent authorities or notices issued by trusted flaggers.[358] For continued content moderation, hosting service providers might have longer periods between the publication of critical content and action.[359]

349 Justin Huang et al., "Politically biased moderation drives echo chamber formation: An analysis of user-driven content removals on Reddit," *SSRN* (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4990476.
350 Tech Against Terrorism, "Patterns of Terrorist Online Exploitation."
351 Tech Against Terrorism, "Patterns of Terrorist Online Exploitation," 7.
352 Tech Against Terrorism, "Patterns of Terrorist Online Exploitation," 9-10.
353 Williams et al., "The Online Extremist Ecosystem," 5-11.
354 Chiara Drolsbach, and Nicolas Pröllochs, "Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database," *WWW '24: Companion Proceedings of the ACM Web Conference* 2024 (2024): 941, https://doi.org/10.1145/3589335.3651482.
355 Drolsbach, and Pröllochs, "Content Moderation on Social Media in the EU," 942.
356 Tarleton Gillespie, "Reduction / Borderline content / Shadowbanning," *Yale-Wikimedia Initiative on Intermediaries & Information* (2022): 2-8, https://law.yale.edu/sites/default/files/area/center/isp/documents/reduction_ispessayseries_jul2022.pdf.
357 Alkiviadou, "Platform liability, hate speech and the fundamental right to free speech," 9-10.
358 See e.g. Art. 9 & 16 DSA. While 'undue delay' is not further defined in the DSA itself, reference to the less than 24 hours response of the Code of Conduct on Countering Illegal Hate Speech Online of 2016 made in § 87 preamble DSA, suggests, that hosting service providers should act on requests and notices under the DSA within one day. Likewise, hosting service providers are required to act on official removal requests within one hour pursuant to Art 3. TCO.
359 See e.g. Art. 5 TCO. This provides that hosting service shall deploy technical mechanisms to 'expeditiously' identify and remove terrorist content, however, without further defining exact timeframes.

Studies examining the automated dissemination of terrorist, illegal, and implicit extremist content online found that flagging – meaning users reporting critical content to the platforms – presents a key aspect of effective content moderation.[360] This is particularly relevant for certain types of platforms, such as gaming platforms, where, due to the interactive and often hidden nature of user communication, other content detection and moderation means are insufficient to address the full scale of terrorist and, in particular, borderline content.[361] However, flagging systems of many major platforms need to be improved, as only ten percent of the content that these researchers flagged on the platforms was eventually removed or disabled after eight weeks.[362] In analysing the activities of AfD-affiliated fake accounts on Facebook, researchers also concluded that the platform's flagging mechanism is insufficient as the relevant forms and mechanisms are not easy to find for users.[363]

6.3.1 Reflecting on Additional Private Sector Initiatives

To tackle cross-platform dissemination[364] of terrorist, illegal or implicit extremist content, hash databases that are accessible to different service providers can be a useful automated tool.[365] These databases contain the characteristics of already identified terrorists or other harmful content. Platforms can then enter these characteristics into their filtering systems to either prevent such content from being uploaded on their platforms or to quickly identify and remove the content.[366]

In another attempt to tackle terrorist content online, several hosting service providers have carried out mass deletions of fake accounts that are known to disseminate terrorist and harmful content.[367] However, research found that such action days, for example, organised by Europol, on which resources are dedicated to delete a large amount of terrorist or harmful content and profiles that spread such content, do not present a sustainable strategy to limit the dissemination of terrorist and harmful content.[368]

Oversight Mechanisms

As the first HSP-led transnational body to adjudicate content moderation decisions, Meta's (previously Facebook) Oversight Board, created in 2019, provides valuable insights into how HSPs can increase legitimacy for their content moderation practices beyond acting as a mere appeals mechanism.[369] At the same time, the work of the Oversight Board points to continuously vague definitions and policies on freedom of speech and the protection of fundamental rights that are being used by platforms in their content moderation.[370] The Board was set up to act as an appeals mechanism for content moderation decisions, brought before it by users and

360 Carpani et al., *EU Internet Forum*, 51-55.
361 Galen Lamphere-Englund, and Menso Hartgers, *CTRL+ALT+COLLABORATE: Public-Private Partnerships to Prevent Extremism in Gaming* (Luxembourg: Publications Office of the EU, 2024), 36-40, https://home-affairs.ec.europa.eu/document/download/e446f013-34e1-4f74-bce6-90f661937ce9_en; Menso Hartgers, and Eviane Leidig, "Fighting extremism in gaming platforms: a set of design principles to develop comprehensive P/CVE strategies," *ICCT* (June 2023), https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it.
362 Carpani et al., *EU Internet Forum*, 43.
363 Albrecht, and Strunk, „Memes für die Massen," 176.
364 Cross-platform dissemination refers to the practice of disseminating the same or similar content across different platforms simultaneously.
365 Galli et al., "The Regulation of Content Moderation," 68-69.
366 One such database for terrorist and violent extremist content is operated by the Global Internet Forum to Counter Terrorism, see Sean Doody, and Michael Jensen, "Hash-Sharing Database Review: Challenges and Opportunities," *GIFCT Year 4 Working Group* (December 2024), https://gifct.org/wp-content/uploads/2025/02/GIFCT-24WG-1224-HSDR-Challenges-1.1.pdf.
367 See e.g. a coordinated effort between Europol and Telegram in 2019. "Europol and Telegram take on terrorist propaganda online," Europol, published November 25, 2019, https://www.europol.europa.eu/media-press/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online.
368 Albrecht, and Strunk, „Memes für die Massen," 156 & 176.
369 For more information on the context and establishment of the Oversight Board, see Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression," *The Yale Law Journal* 129, no. 8 (June 2020), https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3639234.
370 See e.g. Evelyn Douek, "The Meta Oversight Board and the Empty Promise of Legitimacy," *Harvard Journal of Law & Technology* 37, no. 2, (2024): 373 – 445, https://jolt.law.harvard.edu/assets/articlePDFs/v372/3-The-Meta-Oversight-Board-and-the-Empty-Promise-of-Legitimacy.pdf.

the platform itself.[371] Additionally, the "[B]oard can provide policy guidance, specific to a case decision or upon [Meta]'s request, on [Meta]'s content policies." However, this policy advice is non-binding for Meta.[372] Although critics found that the Board is indeed performing these functions, its actual underlying objective is to create legitimacy for Meta's content moderation.[373] At the same time, researchers found that the Board is shying away from clearly outlining the principles of free speech and protection of fundamental rights that Meta and the Board itself are using in conducting content moderation.[374] Another flaw of the Board is the fact that its members do not include experts or practitioners from the tech sector in the process, who would be able to assess the underlying functionalities of Meta's (partially) automated content moderation and its impact on fundamental rights.[375] Nevertheless, in September 2024, the Oversight Board published a non-binding report assessing the use of AI in content moderation by Meta and providing guidance on how to improve these efforts in a human rights compliant manner.[376]

Meta's Oversight Board has also served as an inspiration for other out-of-court dispute settlement and oversight bodies. Notably, in late 2024, the Appeals Centre Europe[377] has been set up with a grant from the Meta Oversight Board Trust with further financing through fees paid by the HSPs that subject themselves to its jurisdiction and users who bring their claims in cases in which decisions is made against them.[378] As of March 2025, Facebook, Instagram, and TikTok allow their users to resort to the Appeals Centre Europe pursuant to Article 21 DSA. However, the Centre does not hear cases on allegedly illegal content but only on content allegedly violating the terms and conditions of the platforms that subjected themselves to its jurisdiction.[379]

Overall, the creation and the work of independent oversight mechanisms have received approval from various governments as well. For example, the Parliamentary Assembly of the Council of Europe in December 2024, called on HSPs to support the work and obey to the decisions of independent oversight mechanisms when conducting content moderation online.[380] Civil society organisations have also long called for the establishment of independent oversight mechanisms to scrutinise the content moderation of HSPs by reviewing moderation decisions, providing guidance on a human rights compliant approach to content moderation, and producing binding decisions for the HSPs under its jurisdiction to improve content moderation practices.[381]

Appeal Procedures

As outlined above, appeal procedures to challenge content moderation decisions present a key obligation under existing EU legal frameworks on terrorist and illegal content online, namely the TCO Regulation and the DSA. It has also been outlined above that appeal procedures are included in most HSPs' content moderation policies to a certain extent. However, in practice, appeals are often insufficient. A study of TikTok's and Instagram's appeal procedures following the deplatforming of users has shown significant hurdles, particularly related to automation of

371 Meta, *Oversight Board Charter* (Menlo Park: Meta, 2019), art 1 (4), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.
372 Meta, *Oversight Board Charter*, art 1 (4).
373 Douek, "The Meta Oversight Board and the Empty Promise of Legitimacy," 384.
374 Douek, "The Meta Oversight Board and the Empty Promise of Legitimacy," 379.
375 Douek, "The Meta Oversight Board and the Empty Promise of Legitimacy," 384.
376 Meta Oversight Board, *Content Moderation in a New Era for AI and Automation* (London: Meta Oversight Board, 2024), https://www.oversightboard.com/wp-content/uploads/2024/09/Oversight-Board-Content-Moderation-in-a-New-Era-for-AI-and-Automation-September-2024.pdf.
377 "Startpagina," Appeals Centre Europe, accessed March 11, 2025, https://www.appealscentre.eu/.
378 Valentina Grippo, *Regulating content moderation on social media to safeguard freedom of expression* (Strasbourg: Parliamentary Assembly of the Council of Europe, 2024), para. 58-59, https://rm.coe.int/as-cult-regulating-content-moderation-on-social-media-to-safeguard-fre/1680b2b162.
379 "Rules of Procedure," Appeals Centre Europe, accessed March 11, 2025, art. 4, https://www.appealscentre.eu/rules-of-procedure/.
380 Grippo, *Regulating content moderation on social media to safeguard freedom of expression*, para. 18.13.
381 ARTICLE 19, *Contents Social Media Councils: One piece in the puzzle of content moderation* (London: ARTICLE 19, 2021), https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf.

these procedures.[382] Notably, affected users felt a lack of information on the reasoning behind their deplatforming as well as on the available appeal mechanisms. They also stated that most of the procedures were automated with little human intervention, which ultimately created the perception of being treated unfairly.[383] This research concluded that while the development of automated tools mainly dealt with content detection and moderation tools, the automated tools in appeal procedures are unjust, creating an overall perception of unfair content moderation and creating high hurdles to challenge and overturn moderation decisions.[384] Nevertheless, experts interviewed for this research acknowledged that AI tools could generally be used to increase transparency and help users understand moderation decisions, given that they provide more accurate information on underlying policies and rationales.[385]

6.4 Role of Regulatory Bodies

The volume of harmful online content is undoubtedly immense, and combined with the concerns that HSPs are struggling to moderate content in a human rights-compliant manner, governmental bodies are increasingly playing a bigger role in content detection. Whether through hard or soft law, these regulatory bodies attempt to overcome some of the shortcomings of content moderation by HSPs.[386] One of their main objectives is to enhance transparency and create more consistency in content detection and moderation. In addition, regulatory bodies have a range of enforcement mechanisms at their disposal, such as ordering mandatory removal of content within a specific timeframe, issuing fines in case of failures to comply with such orders or other administrative sanctions in case of structural shortcomings. The risk of government-led detection and subsequent moderation is that authoritarian regimes or political parties could restrict political dissent.[387] Ensuring that regulatory bodies can function independently is vital to preventing potential political influence in content moderation.

While it goes beyond the scope of this study to elaborate on how the different regulatory bodies at national and European levels operate, it suffices to say that their mandate, approach and functioning differ. The Global Online Safety Regulators Network brings together independent regulators.[388] So far, only nine countries are members of this network, but through sharing experiences, gaps and good practices could be identified.

6.5 Reliability of Detection and Proportionality of Moderation Decisions

While HSPs remain at the forefront of conducting content moderation under the TCO and the DSA, regulators such as the European Commission, or potentially national or intergovernmental enforcement authorities, are taking a more active role in ensuring that all HSPs are indeed willing and able to conduct rule-of-law-based content moderation.[389]

Assessing existing automated content moderation tools that are being used by many HSPs to comply with existing regulations, a considerable body of research emerged, dedicated to

382 Carolina Are, "'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation," *Information, Communication & Society* (2024): 13-14, https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2396621.
383 Are, "'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation," 13-14.
384 Are, "'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation," 15-16.
385 Interview conducted on 11 June 2025, on file with research team.
386 Beatriz Kira, "Regulatory intermediaries in content moderation," *Internet Policy Review* 14, no. 1 (2025), https://policyreview.info/articles/analysis/regulatory-intermediaries-content-moderation.
387 Janjira Sombatpoonsiri, and Sangeeta Mahapatra, "Regulation or Repression? Government Influence on Political Content Moderation in India and Thailand," *Carnegie Endowment for International Peace* (July 2024), https://carnegieendowment.org/research/2024/07/india-thailand-social-media-moderation?lang=en.
388 "The Global Online Safety Regulators Network," eSafety Commissioner, accessed September 10, 2025, https://www.esafety.gov.au/about-us/consultation-cooperation/international-engagement/the-global-online-safety-regulators-network.
389 Watkin, "Developing a Responsive Regulatory Approach to Online Terrorist Content on Tech Platforms," 15.

scrutinise and improve existing tools, and to encourage the development of new detection and moderation tools.[390] Overall, research calls for the regulation of AI-driven content moderation tools to ensure harmonised application across platforms and consistent protection of users' fundamental rights.[391]

The trustworthiness and legitimacy of algorithms and AI tools in content detection and moderation remain uncertain. One limitation of current research is its focus on account and content removal, despite platforms taking various, often less transparent, moderation actions, such as lowering the visibility of content or adding flags and notes.[392] Disclosing all aspects of platforms' content moderation options appears essential to understanding the varying risks posed to fundamental rights in the online environment.[393] To ensure the accuracy and impartiality of algorithmic decision-making, these systems must be accompanied by transparent logs that clearly outline the rationale behind their decisions. However, this push for transparency may inadvertently lead to over-enforcement, creating a more censored online sphere.[394] So far, the Dutch cases related to the detection and moderation of online content—such as the Danny Mekić v. X (Twitter) court decision[395]—were primarily focused on platform transparency obligations, including algorithmic moderation, shadow banning, and user redress, and not specifically on terrorist content, illegal content or implicit extremist content.

6.6 False Positives: Impact on the Right to Freedom of Expression Online

In today's world, human rights violations are increasingly committed by non-state actors, including armed groups, corporations, and even individuals within communities. While international human rights law primarily focuses on state responsibility, there's a growing recognition that non-state actors must also be held accountable for their actions that impact human rights.[396] This also relates to how content moderation can infringe on the freedom of expression, but also other human rights can be impacted by decisions of HSPs. For example, data collection can affect the right to privacy of users, and biased algorithms can lead to discrimination.

Reports indicate that, to date, HSPs have failed to achieve an appropriate balance between safeguarding freedom of expression and enforcing content moderation policies.[397] The right to freedom of expression appears critically threatened online, with a discernible decline in public discourse. HSPs often implement vague operational guidelines, which can result in censorship without adequate scrutiny of the content in question.[398] Additionally, content moderation, if not carefully managed, can disproportionately affect the right to freedom of expression of minority, marginalised groups, or activist groups when content targeting these groups is removed. For instance, content moderation mechanisms may include biases related to sexual orientation, race,

390 In addition to pieces already cited in this report, see e.g. Szu-Yin Lin et al., "Combating Online Malicious Behavior: Integrating Machine Learning and Deep Learning Methods for Harmful News and Toxic Comments," *Information Systems Frontiers* (2024), https://doi.org/10.1007/s10796-024-10540-8.
391 Althaf Marsoof et al., "Content-filtering AI systems–limitations, challenges and regulatory approaches," *Information & Communications Technology Law* 32, no. 1 (2023): 64-101, https://doi.org/10.1080/13600834.2022.2078395.
392 Arora et al., "Detecting Harmful Content on Online Platforms," 9.
393 EU Agency for Fundamental Rights, *Online content moderation - Current challenges in detecting hate speech* (Luxembourg: Publications Office of the EU, 2023), 11, https://fra.europa.eu/en/publication/2023/online-content-moderation.
394 Giancarlo Frosio, "Algorithmic Enforcement Tools: Governing Opacity with Due Process," in *Driving Forensic Innovation in the 21st Century* (Cham: Springer, 2024), 204, https://doi.org/10.1007/978-3-031-56556-4_9.
395 Danny Mekić v. Twitter (X), 742407 / HA RK 23-366 Rechtbank Amsterdam (2024), https://uitspraken.rechtspraak.nl/\details?id=ECLI:NL:RBAMS:2024:4019&showbutton=true&keyword=C%252f13%252f742407&idx=1.
396 Rikke Frank Jørgensen, "When private actors govern human rights," in *Research Handbook on Human Rights and Digital Technology* (Cheltenham: Edward Elgar Publishing, 2019), 346-363.
397 ARTICLE 19, *Content moderation and freedom of expression handbook*, 46.
398 "Article 12 DSA: Will platforms be required to apply EU fundamental rights in content moderation decisions?" Naomi Appelman et al., DSA Observatory, published December 9, 2024, https://dsa-observatory.eu/2021/05/31/article-12-dsa-will-platforms-be-required-to-apply-eu-fundamental-rights-in-content-moderation-decisions/.

gender, or religion, which could lead to discrimination against specific groups.[399] As a result, users from these targeted groups may hesitate to share their views on social media for fear of having their content removed.

Online research and expert consultation carried out for this study indicate that false positives are a result of numerous factors, notably:

- Wrongfully classifying content due to vague definitions;
- Rightly classifying content, but disproportionally moderating the content;
- Lack of procedural safeguards, such as a lack of notification or appropriate appeal mechanisms to challenge the decision.

Considering that HSPs have the ability to shape public opinion, amplify content, including harmful content and impact human rights, they also have a responsibility to protect human rights. Indeed, according to UN Guiding Principles on Business and Human Rights (UNGPs), HSPs have a responsibility to respect human rights, which includes mitigating risks of human rights violations and taking measures to prevent human rights violations.[400] Although comprehensive guidelines have been established at the European and international levels to provide a structured framework for safeguarding fundamental human rights while moderating online content,[401] the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, companies fail to implement a human rights-compliant framework to moderate content online.[402]

Lastly, independent courts of law play a pivotal role in adjudicating cases concerning content moderation and the right to freedom of expression. It is essential for the judiciary to rigorously evaluate whether online content moderation actions infringe upon an individual's fundamental right to freedom of expression.[403] This requires carefully balancing the competing rights and interests involved to ensure a thorough and equitable review of content moderation decisions.[404] Nevertheless, domestic courts seem to not always achieve this objective, as evidenced by the numerous cases brought before the ECtHR for a final decision on how to balance competing rights on online content moderation.[405] While ECtHR mainly looks at whether States have unlawfully restricted the freedom of expression and not HSPs, some of the indicators used by the court in cases related to the freedom of expression in general could be helpful for HSPs. The analysis of these indicators, unfortunately, falls outside the scope of the current research. Furthermore, legal proceedings, if they are pursued by affected stakeholders in the first place, often take considerable time, which may delay judicial guidance on effective protection of the right to freedom of expression. Once decisions are rendered, then HSPs can align their policies and practices with judicial decisions, however, not all decisions are binding on private actors but rather binding for states to ensure proper safeguards.

Scholars have identified important challenges regarding the respect of freedom of expression while countering illegal and harmful content. Notably, when algorithms are employed to detect

399 Wolfgang Schulz et al., *Algorithms and human rights - Study on the human rights dimensions of automated data processing techniques and possible regulatory implications* (Strasbourg: Council of Europe, 2018), 27-28, https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html.
400 UN Human Rights Council, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (New York & Geneva: UN, 2011), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
401 European External Action Service, *EU Guidelines on Freedom of Expression Online and Offline* (Brussels: Council of the EU, 2018), 1, https://www.eeas.europa.eu/sites/default/files/09_hr_guidelines_expression_en.pdf.
402 David Kaye, *Promotion and protection of the right to freedom of opinion and expression* (New York: UN, 2019), 16, https://docs.un.org/en/A/74/486.
403 Erik Tuchtfeld, "Case law on content moderation and freedom of expression," *The Global Freedom of Expression Special Collection of the Case Law on Freedom of Expression* (June 2023): 8, https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2023/06/GFoE_Content-Moderation.pdf.
404 Tuchtfeld, "Case law on content moderation and freedom of expression," 8.
405 European Court of Human Rights, *Factsheet - Hate speech* (Strasbourg: Council of Europe, 2023), https://www.echr.coe.int/documents/d/echr/FS_Hate_speech_ENG.

and moderate such content, these practices may inherit biases and lack transparency, accuracy, and accountability.[406] The TCO as well only considers the rights of individuals after a removal order is issued, while its main impact stems from proactive measures, as HSPs see themselves compelled (for financial and economic reasons) to act before a removal order is received in light of the financial penalties they could face, as outlined above. This ties into proactive and reactive approaches to content moderation. The automatic use of filters and hash databases prevents certain harmful content from even being posted, but if content is falsely classified, it can infringe on freedom of expression. As a result, HSPs may interpret the TCO Regulation expansively, including its broad terrorism definition, when creating content moderation policies to avoid sanctions.[407] Additionally, the ability to issue removal orders without judicial oversight raises concerns about potential censorship and calls for the establishment of appropriate safeguards, tailored to the risks and methods of enforcing cross-border removal orders.[408] On the other hand, the creation of a new obligation for VLOPs and very large online search engines (VLOSEs) to conduct fundamental rights impact assessments, as outlined in Article 34(1)(b) of the DSA, was welcomed by scholars.[409] This requirement ensures that the protection of fundamental rights is integrated into the systems, processes, and policies of VLOPs and VLOSEs, thereby providing enhanced protection for users.[410] It remains to be seen how the VLOPs and VLOSEs will implement this obligation.

In light of the various challenges faced in moderating online content in accordance with human rights principles, the UK Independent Reviewer of Terrorism Legislation argues that the application of fundamental human rights language in the online context may be inadequate or require modification to appropriately consider the broader public interest, especially when the likelihood of terrorist content inciting violence is minimal. He asserts that an alternative approach is necessary, one that prioritises greater transparency regarding the trade-offs involved, to more effectively balance the competing interests at hand.[411] In a 2020 study, the Council of Europe evaluated how the various stakeholders involved in the moderation of hate speech could benefit from collaborative efforts, such as closer alignment with human rights advocates representing the interests of online users, to foster the adoption of a more human rights-based approach in addressing this issue.[412] Civil rights experts interviewed for this study also confirmed positive experiences in collaboration between civil society and online platforms to develop human-rights-based policies for the use of AI in content detection and moderation.[413]

Ultimately, balancing freedom of speech with the need to counter terrorism online lies at the heart of developing a human rights-based approach to terrorist and implicit extremist content online. Beyond the elaborations above, this is also illustrated by the fact that white supremacists and right-wing extremists move to alternative platforms, including encrypted messenger

406 Felipe Romero-Moreno, "Generative AI and deepfakes: a human rights approach to tackling harmful content," *International Review of Law, Computers & Technology* 38, no. 3 (2024): 299, https://doi.org/10.1080/13600869.2024.2324540.
407 Tarik Gherbaoui, and Martin Scheinin, "A Dual Challenge to Human Rights Law: Online Terrorist Content and Governmental Orders to Remove it," *Journal européen des droits de l'homme - European Journal of Human Rights* (2023): 7, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4247120.
408 Marcin Rojszczak, "Gone in 60 Minutes: Distribution of Terrorist Content and Free Speech in the European Union," *Democracy and Security* 20, no. 2 (2023): 193-197, https://www.tandfonline.com/doi/abs/10.1080/17419166.2023.2250731.
409 Eliška Pírková et al., *Towards Meaningful Fundamental Rights Impact Assessments Under The DSA* (The Hague & Brussels: European Center for Not-for-Profit Law & Access Now Europe, 2023), 33, https://ecnl.org/sites/default/files/2023-09/Towards%20Meaningful%20FRIAs%20under%20the%20DSA_ECNL%20Access%20Now.pdf.
410 Pírková et al., *Towards Meaningful Fundamental Rights Impact Assessments Under The DSA*, 7.
411 Jonathan Hall, "Rights and Values in Counter-Terrorism Online," *Studies in Conflict & Terrorism* (2023): 9, https://www.tandfonline.com/doi/full/10.1080/1057610X.2023.2222889.
412 Alexander Brown, *Models of Governance of Online Hate Speech: On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe* (Strasbourg: Council of Europe, 2020), 26-28, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d.
413 Interview conducted on 11 June 2025, on file with research team.

services, when their content gets moderated on other platforms.[414] Extensive removal of online content also impedes the ability to identify key individuals, enabling extremists to continue their activities and network without interference. Hence, content moderation efforts should always be accompanied by cooperation with law enforcement and prosecution to ensure that harmful content is not only removed online in a rule-of-law-based manner but can also be used in holding individuals accountable and addressing underlying structures.

6.7 Gaps and Observations

Researchers have for several years pointed out that, especially in light of diffusing ideologies, content moderation must pay more attention to so-called borderline content as it could contribute to the proliferation of extremist views online.[415]

Determining the lawfulness of content is becoming increasingly complex. In addition to the plethora of regulation, the volume of online content, the increasing use of AI and concealing tactics that are being used, HSPs face difficulties in effectively detecting and moderating content. As a result, HSPs often fail to identify certain content as hate speech or other forms of illegal content.[416] A particular concern arises when users employ implicit language or memes understood only by specific groups or communities, thereby evading detection by content moderation systems, despite the content being harmful or unlawful.[417] As this content remains visible online, it can reach a broader audience and support the growth of harmful communities. The right to freedom of expression takes on a different dimension online. The notion of 'free speech is not the same as free reach' referring to the fact that you may have the freedom to express your opinions, but not the right to have it widely promoted, is being contested. Platforms, albeit being private companies, control the possibilities to amplify, demote, or reduce the visibility of content.[418]

As both the TCO and DSA have only recently entered into force, it remains difficult to assess what the main challenges and benefits of these regulations are and whether their interpretation by HSPs may lead to too many false positives or false negatives. This was also confirmed by several experts interviewed for this study.[419]

The terms of use, community guidelines and policies of platforms are not easily accessible, making it difficult for users to find out what is permissible by platforms and how to challenge moderation decisions. Furthermore, this lack of transparency regarding detection and moderation has resulted in limited research on users' perceptions and experiences with these actions, notably regarding the fairness of these tools.[420] However, an experiment revealed that individuals who were notified of their violation of platform community standards, accompanied by a link to the relevant rules, were more likely to comply with those standards.[421] Other researchers have highlighted the potential of AI tools to support moderators in making complex contextual judgments, especially when classification-based tools in content moderation are primarily used

414 Williams et al., "The Online Extremist Ecosystem."; Alexandra Siegel, "Online Hate Speech," in *Social Media and Democracy: The State of the Field and Prospects for Reform* (Cambridge: Cambridge University Press, 2020), 72, https://doi.org/10.1017/9781108890960; Aleksandra Urman, and Stefan Katz, "What they do in the shadows: examining the far-right networks on Telegram," *Information, Communication & Society* 25, no. 7 (2022): 904–923, https://doi.org/10.1080/1369118X.2020.1803946.
415 Won, and Lewis, "Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts."
416 EU Agency for Fundamental Rights, *Online content moderation*, 9.
417 Won, and Lewis, "Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts."
418 Evelyn Douek, "Content moderation as systems thinking," *Harvard Law Review* 136, no. 2 (2022): 545-546, https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf.
419 Interviews conducted on 11 June 2025 and 25 June 2025, on file with research team.
420 Devesh Narayanan et al., "Fairness Perceptions of Artificial Intelligence: A Review and Path Forward," *International Journal of Human–Computer* Interaction 40, no. 1 (2024): 5-6, https://doi.org/10.1080/10447318.2023.2210890.
421 Matthew Katsaros et al., "Online Content Moderation: Does Justice Need a Human Face?" *International Journal of Human–Computer Interaction* 40, no. 1 (2024): 70, https://doi.org/10.1080/10447318.2023.2210879.

to flag content for human review.[422] Although limited research has focused on the application of machine learning in detecting radical behaviours and terrorist activities in cyberspace, its successful use in other cyber domains suggests that it could yield similarly effective results in identifying terrorist content online.[423]

Lastly, despite increasing scholarly attention to data protection, research on strategic litigation in the field of digital rights remains limited.[424] Existing studies predominantly examine the opportunities and challenges associated with the enforcement of the GDPR, yet comparatively little attention has been devoted to broader legal mobilisation efforts within the digital sphere.[425] Consequently, a significant gap remains in the analysis of how strategic litigation operates across different domains of the EU digital regulatory framework and the factors influencing its effectiveness.

422 Stuart Macdonald et al., *Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online* (London: Tech Against Terrorism, 2024), 17-18, https://tate.techagainstterrorism.org/news/tcoaireport.
423 Reza Montasari, "Machine Learning and Deep Learning Techniques in Countering Cyberterrorism," in Cyberspace, Cyberterrorism and the International Security in the *Fourth Industrial* Revolution (Cham: Springer, 2024), 136, https://doi.org/10.1007/978-3-031-50454-9_8.
424 Valentina Goluvana, and Sarah Tas, "Guardians of Digital Rights: Exploring Strategic Litigation on Data Protection and Content Moderation in the EU," *Nordic Journal of European Law* 7, no. 4 (2024): 50, https://journals.lub.lu.se/njel/article/view/27306/24043.
425 Goluvana, and Tas, "Guardians of Digital Rights," 50.

7. Feasibility of an Assessment Framework

7.1 Introduction

In this chapter, we shift the focus to the key research question related to the feasibility of setting up an assessment framework for online content. Based on the legal elements of terrorist content and illegal content that follow from the legal frameworks elaborated in chapter 4, and the indicators of implicit extremist content that could be harmful identified in chapter 5, a pilot codebook is presented to assist in detecting the false negatives. After the presentation of the pilot codebook, we reflect on the usability of this pilot codebook. We do this by using our sample of content collected through our open-source intelligence (OSINT) research. Questions that will be addressed relate to detecting and qualifying online content as terrorist, illegal or implicit extremist content and appropriately moderating it. We will elaborate on the observations made during the coding of the research team, and highlight where we identified challenges related to the reliability of the coding of the indicators. We furthermore offer ideas regarding the combination of indicators necessary to identify implicit extremist content.

7.2 Drafting a Pilot Codebook

Based on the desktop research we conducted for chapters 4 and 5, we identified key indicators for the qualification of terrorist, illegal, and implicit extremist content online. Considering the fact that the objective of this study is not the design of an assessment framework in itself, but rather a feasibility test of such a framework, the scope of the pilot codebook is limited. Particularly since the aim was also to do a test run of the pilot codebook by coding the sample of content collected through our OSINT research, and the coding itself is very labour intensive, the number of indicators needed to be limited. The research team therefore did not identify a complete list of indicators for the identification of all forms of terrorist, illegal, and implicit extremist content, but rather limited the number of qualifications related to terrorist, illegal, and implicit extremist content, and henceforth the number of indicators used. For the purpose of this study and to test the feasibility of the pilot codebook, the focus for the selection of the test sample was on right-wing extremist, jihadist and borderline content, meaning content that did not fall clearly in the first two categories because of concealed language or because it was not obviously illegal. The research team selected the type of terrorist content that can be most disputed. Regarding the qualification of illegal content, only a limited number of other types of illegal content have been included to test the feasibility of the framework. These have been selected based on their potential overlap with the category of implicit extremist content.

7.2.1 Pilot Codebook

The pilot codebook uses a step-by-step approach to qualify content as terrorist content, illegal content, and implicit extremist content. It consists of four steps. For steps 1 and 2, the qualification of content as either terrorist content or illegal content means that once identified as such, the coding of content would end. If such is the case, the assumption is that the detection of terrorist content or illegal content is followed by a moderation decision related to the content.

If, after step 1, no terrorist content is detected, the coding continues with step 2. The same procedure is followed after steps 2 and 3. When in step 3, indicators are present to suggest that one is dealing with implicit extremist content, the content needs to be scored again in accordance with steps 1 and 2, to consider whether, after the concealing tactics used to conceal true meaning are revealed, the content could, at second glance nevertheless qualify as terrorist or illegal content or as lawful but awful.

- Step 1: Assess whether it contains terrorist content (TC)
  - If the content qualifies as terrorist content ➔ stop coding
  - If it does not qualify as terrorist content ➔ proceed to step 2

- Step 2: Assess whether it contains other explicit illegal content (EIC)
  - If content qualifies as explicit illegal content ➔ stop coding
  - If it does not qualify as ➔ proceed to step 3

- Step 3: Assess whether it contains implicit extremist content that is potentially harmful (IHC)
  - If content does not qualify as implicit extremist content, it falls under the right to freedom of expression ➔ stop coding
  - If it does qualify as implicit extremist content ➔ proceed to step 4

- Step 4: Assess whether this implicit extremist content that is potentially harmful is implicitly illegal
  - If the implicit extremist content is legal ➔ content is considered 'harmful but lawful' and falls under the right to freedom of expression
  - If the implicit extremist content is illegal ➔ content is considered 'harmful unlawful'

The pilot codebook is set up in Excel (see the clean template included in the Annex 7) and allows for each of the 320 posts collected during the OSINT research phase to register general information about the post, before scoring the indicators relevant in the different steps. The coding method simply registers 0 if the content does not contain the information identified by the indicator, or 1 if the content does contain that information. Below, a detailed description of the general information registered and the indicators included in each step of the pilot codebook is provided.

7.2.2 Codebook Indicators and Scoring Steps

Step 1 – Is it Terrorist Content (TC)?

This first step aims to identify terrorist content among social media posts collected as part of our OSINT research. The indicators used to determine whether a piece of content is terrorist content are based on TCO Regulation (EU Regulation 2021/784) and thus are binding for all HSPs falling under the TCO as well as the competent authorities in the EU MS and at the EU level.

The scores/variations for each indicator correspond to the elements of the crime, as well as a last score/variation ensuring that the content does not fall under freedom of expression. All scores/variations have to be met for the content to fall under one of the terrorist content types.

Table 3: Indicators for terrorist content

Indicator/variable	Scores/variations
Incitement to commit/ participate in a terrorist offence (IT) *Soliciting a person or a group of persons to commit or contribute to the commission of a terrorist offence*	The intended audience is a group
	The audience is sufficiently clear to include people with certain ideological ties
	The content is spread in public
	Advocating the commission of a terrorist offence
	Awareness of the likelihood that audience member(s) could commit a terrorist act (**capacity of audience**)
	Deliberately creating the likelihood that a **serious** terrorist offence will be committed regardless of whether it will eventually be executed or not
	There is a direct or indirect **causal link** between incitement and terrorist act
	There is an **imminent** risk that a terrorist act will be committed
	Knowledge of or reference to terrorist activities by a group or individual by reference to either: participation in a terrorist group/recruitment activities for a terrorist group/providing training for a terrorist group/receiving training for a terrorist group/travelling for the purpose of terrorism/organising or facilitating travel for terrorism/terrorist financing/extortion with a view to commit a terrorist offence/ creating or using fake document to commit a terrorist offence/aggravated theft to commit a terrorist offence.[426]
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation
Glorification of a terrorist act (GT) *Directly or indirectly justifying or advocating the commission of terrorist offences*	Justification of a past or future serious terrorist offence/advocating the commission of a serious terrorist offence
	Deliberately creating an atmosphere of fear
	Knowledge of terrorist activities by reference to either: participation in a terrorist group/recruitment activities for a terrorist group/providing training for a terrorist group/receiving training for a terrorist group/travelling for the purpose of terrorism/organising or facilitating travel for terrorism/terrorist financing/extortion with a view to commit a terrorist offence/ creating or using face document to commit a terrorist offence/aggravated theft to commit a terrorist offence[427]
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation

426 This variable is meant to assess the individual criminal responsibility, meaning whether a person acted with intent and/or knowledge of the crime (mens rea). This variable is required in order to examine whether the individual who created or disseminated certain content meant to engage in the alleged conduct, meant to cause a certain consequence or was aware that this consequence will likely occur, and knew of this consequence. This step is being taken to avoid flagging potentially harmful content which was created and/or disseminated without malign intentions and to exclude person who for example have a lawful justification such as been forced or are mentally incapable. At the same time, researchers acknowledge that this is merely a preliminary superficial assessment.

427 This variable is meant to assess the individual criminal responsibility, meaning whether a person acted with intent and/or knowledge of the crime (mens rea). This variable is required in order to examine whether the individual who created or disseminated certain content meant to engage in the alleged conduct, meant to cause a certain consequence or was aware that this consequence will likely occur, and knew of this consequence. This step is being taken to avoid flagging potentially harmful content which was created and/or disseminated without malign intentions and to exclude person who for example have a lawful justification such as been forced or are mentally incapable. At the same time, researchers acknowledge that this is merely a preliminary superficial assessment.

Recruitment for a terrorist organisation (RT) *Soliciting a person or a group of persons to participate in the activities of a terrorist group*	Directly or indirectly addressing an audience (individual or group)
	Call to supply information/material resources/funding/human resources to a terrorist group
	Knowledge of or reference to a specific terrorist group, one or more serious criminal acts committed by the referenced terrorist group, one or more concrete aims of the terrorist group
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation
Indicators not included that should be considered for a comprehensive framework: • Threat to commit a terrorist offence as provided in Art. 3 (1) lit. a-I EU Directive 2017/541 • Instructions on the making and use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing a terrorist offence • Terrorist financing	

Coding process:

- If Yes (1) to one terrorist content (TC) indicator ➔ The content might qualify as terrorist content ➔ Stop coding
- If No (0) to all terrorist content (TC) indicators ➔ Continue to Step 2

Step 2 – Is it Explicitly Illegal Content (EIC)?

Social media posts identified as potential terrorist content have been filtered through Step 1. This second step focuses on detecting other forms of illegal content (i.e. non-terrorist) among the remaining posts. For the purpose of this pilot codebook, only a limited number of other illegal content types have been included to test the feasibility of the framework. As mentioned before, these have been selected by the research team based on their potential overlap with the category of implicit extremist content.

For each indicator, the scores/variations are based on the elements of the respective crimes as codified in the Dutch Criminal Code, as well as a last score ensuring that the content does not fall under freedom of expression. All scores/variations have to be met for the content to fall under one of the terrorist content types.

Table 4: Indicators for illegal content

Indicator/variable	Scores/variations
Hate Speech (HS)	The content is spread in public
	The content discriminates a group or an individual
	The content discriminates on the basis of race, gender, religion or belief, sexual orientation, physical or mental disability
	The content insults/incites hatred or violence
	Knowledge of or reference to discriminatory acts or speech[428]
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation
Incitement to Violence (IV)	The content is spread in public
	There is a causal (direct and imminent) link between incitement and serious violent offence
	Advocating the commission of a serious violent offence
	Awareness of the likelihood that audience member(s) could commit a serious violent offence
	Deliberately creating the likelihood that a serious violent offence will be committed regardless of whether it will eventually be executed or not
	The intended audience is a group
	Knowledge of or reference to serious acts of violence[429]
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation
Denial, downplaying or justification of international crimes (DIC)	The content is spread in public
	It is against a group or an individual
	The person denies, justifies, or downplays core international crimes that have been established irrevocably by a court
	Knowledge of or reference to the fact that crimes constitute core international crimes[430]
	Content is **not** satire/parody/artistic expression/legitimate contribution to the public discourse/legitimate commemoration of historical events, colonial past or decolonialisation

428 This variable is meant to assess the individual criminal responsibility, meaning whether a person acted with intent and/or knowledge of the crime (mens rea). This variable is required in order to examine whether the individual who created or disseminated certain content meant to engage in the alleged conduct, meant to cause a certain consequence or was aware that this consequence will likely occur, and knew of this consequence. This step is being taken to avoid flagging potentially harmful content which was created and/or disseminated without malign intentions and to exclude person who for example have a lawful justification such as been forced or are mentally incapable. At the same time, researchers acknowledge that this is merely a preliminary superficial assessment.
429 This variable is meant to assess the individual criminal responsibility, meaning whether a person acted with intent and/or knowledge of the crime (mens rea). This variable is required in order to examine whether the individual who created or disseminated certain content meant to engage in the alleged conduct, meant to cause a certain consequence or was aware that this consequence will likely occur, and knew of this consequence. This step is being taken to avoid flagging potentially harmful content which was created and/or disseminated without malign intentions and to exclude person who for example have a lawful justification such as been forced or are mentally incapable. At the same time, researchers acknowledge that this is merely a preliminary superficial assessment.
430 This variable is meant to assess the individual criminal responsibility, meaning whether a person acted with intent and/or knowledge of the crime (mens rea). This variable is required in order to examine whether the individual who created or disseminated certain content meant to engage in the alleged conduct, meant to cause a certain consequence or was aware that this consequence will likely occur, and knew of this consequence. This step is being taken to avoid flagging potentially harmful content which was created and/or disseminated without malign intentions and to exclude person who for example have a lawful justification such as been forced or are mentally incapable. At the same time, researchers acknowledge that this is merely a preliminary superficial assessment.

Indicators not included in this blueprint that should be considered for a comprehensive framework:
- Defamation
- Child sexual abuse material
- Weapons offences

Coding process:
- If Yes (1) to one of the explicit illegal content (EIC) indicators ➜ This content might qualify as illegal content ➜ Stop coding
- If No (0) to all explicit illegal content (EIC) indicators ➜ Continue to Step 3

Step 3 – Is it Implicit Extremist Content that is potentially harmful (IHC)?

Through Steps 1 and 2, social media posts which explicitly appear to fall into the categories of terrorist or other forms of illegal content have been filtered. The objective of this third step is to identify the presence of implicit extremist content that is harmful, which may not immediately appear as illegal, among the remaining posts' 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.

Implicit refers to the fact that the content is concealed, and when this is done intentionally, it aims to disguise the illegality, unlawfulness or harmfulness of the content. The step thus aims to lift the veil and uncover the true meaning of the content. In order to do so, the research team has identified – building upon preliminary observations made during the data collection phase, existing research consulted for the literature review, and the team's own expertise in the field – a limited number of indicators to test the feasibility of a framework. These indicators have been selected based on their susceptibility to reveal (evasive techniques used to hide) the implicit harmfulness of the content, as follows:

- The **intent to conceal meaning (CM)** indicator captures elements that suggest an effort to obscure the 'true' meaning of the narrative put forward. This indicator and its scores/variations have been identified based on the various textual and visual manipulation techniques observed during the data collection phase.
- The **harmful alliances/affiliations (HA)** indicator focuses on all signs denoting a connection to hateful or violent extremist groups or ideologies. This indicator and its scores/variations have been identified based on both preliminary observations made during the data collection phase and existing open-source databases documenting hateful and extremist symbols, emojis, and other coded language.[431]
- The **problematic reference(s) to the historical/current context (PR)** indicator aims to situate the content within a broader context and historical narratives. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).
- The **implicit action trigger(s) (AT)** indicator identifies elements inserted in the narratives that might subtly encourage the audience to take actions. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).
- The **presumed intent to cause harm (IH)** indicator considers potential harmful intentions

431 See e.g. ADL, *Hate On Display™: Hate Symbols Database* (New York: ADL, 2019), https://www.adl.org/sites/default/files/ADL%20Hate%20on%20Display%20Printable_0.pdf; "Global Extremist Symbols Database," Global Project Against Hate and Extremism, accessed September 10, 2025, https://globalextremism.org/global-extremist-symbols-database/; "Interpreting and Translating Emojis," International Centre for Digital Threat Assessment, *Interpreting and Translating Emojis* (Surrey: Safer Schools Together, 2025), https://resources.saferschoolstogether.com/view/294238754/.

behind content, focusing on *presumed* rather than definitive intent due to the inherent subjectivity in such assessments. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

Contrary to the previous steps, only one score/variation needs to be present to consider that the related indicator is present (e.g. if the content contains hateful or extremist symbols, but no emojis or slogans, the harmful alliances/affiliations indicator will be considered present). Whether the presence of one indicator in the post is enough to qualify the content as implicit extremist content is related to the question of the threshold that needs to be met (see below).

Table 5: Indicators for implicit extremist content

Indicator/variable	Scores/variations
Intent to conceal meaning (CM) *Elements in the content translate an intent to conceal its harmful nature.*	Ambiguous or coded language
	Altered text or misspelling
	Misleading cover image (video)
	Blurred or altered image (e.g. marker function to hide certain words)
	Use of humour or irony
	Other (specify)
Harmful alliances/ affiliations (HA)* *Signs denoting affiliation to a hateful / violent extremist group or ideology.*	Hateful / extremist symbols (RWX)
	Hateful / extremist symbols (Islamist)
	Hateful / extremist emojis (RWX)
	Hateful / extremist emojis (Islamist)
	Hateful / extremist coded slogans, slang, or acronyms (RWX)
	Hateful / extremist coded slogans, slang, or acronyms (Islamist)
	Other (specify)
Problematic reference(s) to historical/current context (PR) *Harmful (re-) interpretations, instrumentalisation, disinformation about historical and/or current events.*	Denial or questioning of proven past or present crimes (e.g. dismissal of judicial rulings, disputing evidence, non-recognition of victims, minimisation of victimhood, etc.)
	False or misleading claims about past or present crimes (e.g. crimes that never occurred, crimes acquitted through judicial ruling, misattribution of responsibility, misattribution of victimhood, etc.)
	Justification of current or potential future crimes through references to past crimes, whether real or alleged
	Falsified historical claims aimed at denying the existence or territorial legitimacy of a state
	Glorification or positive portrayal of individuals or groups involved in crimes (e.g. war criminals, 'martyrs', collaborators, etc.)
	Glorification or positive portrayal of public figures known for spreading antisemitic, xenophobic, Islamophobic, and otherwise hateful or extremist narratives (e.g. politicians, scholars, religious figures, historians, etc.)
	Promotion or endorsement of books, films, essays, or any other products known for spreading or supporting any of the above narratives
	Other (specify)

Implicit action trigger(s) (AT) *Elements in the content that might implicitly trigger the audience into taking specific actions by...*	Instilling a perceived need for retaliation
	Instilling a perceived need for self-defence
	Instilling a perceived need to protect a group from reputational damage
	Instilling the idea of a compensation / reward for undertaking action
	Exerting peer pressure / respect for a code of honour
	Providing sacred justification for action
	Other (specify)
Presumed Intent to cause harm (IH) *Elements in the content denote an intent to...*	Normalise hateful / violent narratives
	Propagate hateful / violent conspiracies
	Foster hate / hostility towards an out-group
	Trigger to take online harmful actions (e.g. harassment/doxxing)
	Trigger to take offline harmful actions
	Other (specify)
These indicators have been selected based on the case studies for their susceptibility to yield results in terms of implicit extremist content related to violent extremism. It excludes other forms of harmful content, such as: • Animal cruelty • Self-harm	

In order to help assess what threshold should be applied at Step 3 to determine whether content may be considered as implicitly extremist and requiring further examination under Step 4, the research team has integrated into the Excel Codebook a few options that will be tested through the coding:

Figure 2: Post #107

1. **Single indicator:** One option could be to consider that if any one of the listed indicators is present, the content would automatically qualify as implicitly harmful and proceed to Step 4 for further examination. *Option 1: If Yes to one IHC indicator → Proceed to Step 4.*

However, the research team's questions would strongly oppose this approach. Since the use of humour is one of the indicators to conceal the meaning of the post, ticking this box, for instance, without assessing whether a specific problematic meaning is at all concealed, would make no sense. See, for instance, post #107, clearly involves humour, and although the meaning is not fully clear, there is no further problematic reference made in the post.

2. **Multiple indicators:** Another option could be to consider that no single indicator alone can determine whether content is implicitly harmful. In this case, a certain number of indicators would have to be present for the content to qualify as implicitly harmful and requiring further examination under Step 4. *Option 2: If Yes to at least 2 IHC indicator ➔ Proceed to Step 4*

3. **Combined indicators:** A last option could be to consider that certain indicators are measuring different aspects of the content (e.g. implicitness vs. harmfulness). Under this approach, content would proceed to Step 4 only if a certain combination of indicators is present.*Option 3a: If Yes to "Intent to Conceal Meaning (CM)" and 1 other IHC indicator ➔ Proceed to Step 4 Option 3b: If Yes to "Presumed Intent to Cause Harm (IH)" and 1 other IHC indicator ➔ Proceed to Step 4*

Coding process:

- If No (0) to *[threshold TBC]* ➔ This content does not appear to be extremist and falls under the right to freedom of expression à Stop coding
- If Yes (1) to *[threshold TBC]* ➔ This content might be extremist, even if implicit, and requires further examination to determine whether it is legal or illegal ➔ Continue to Step 4

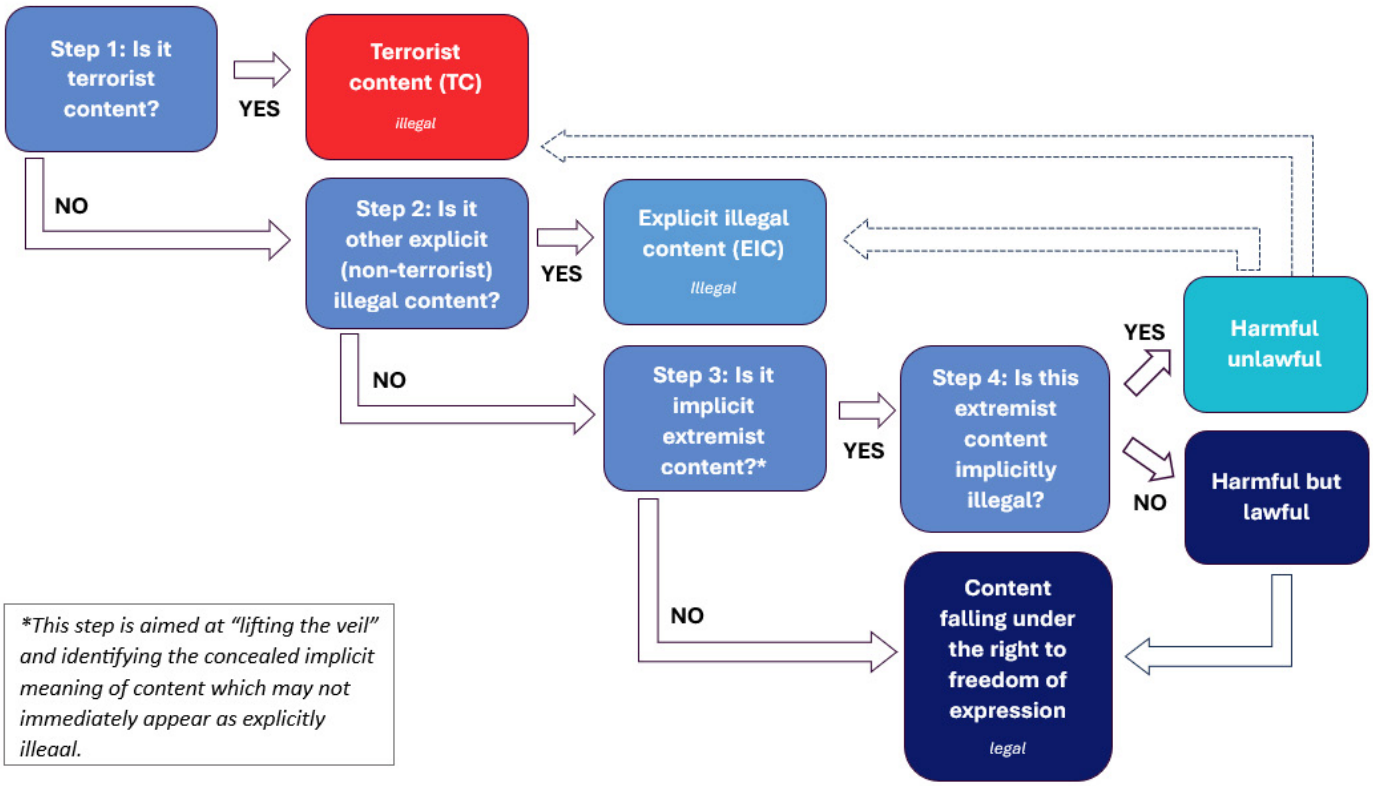
Step 4 (Feedback Loop) – Is this implicit extremist content implicitly illegal?

Step 3 has allowed to lift the veil and uncover the implicit meaning and harmfulness of a certain piece of content. While this content may not have immediately appeared as explicitly terrorist or illegal earlier in the process (i.e. Steps 1 & 2), it is necessary to re-consider its legality based on the uncovered elements identified in Step 3 (e.g. coded language, hate symbols, emojis, etc.). For example, what may seem like innocuous content at first glance could, upon closer analysis, be found to glorify terrorist acts or spread hate speech through hidden or coded messaging. Step 4 is aimed at allowing for this re-assessment of the legality of implicit content.

To this end, it introduces a feedback loop in the model that prompts the assessor to return to Steps 1 and 2 in order to re-evaluate whether the content – now understood to contain a coded or concealed message – falls within the scope of terrorist or other forms of illegal content.

- If Yes (1) to 1 TC indicator in Step 1 ➔ The content is terrorist content
- If Yes (1) to 1 EIC indicator in Step 2 ➔ The content is explicit illegal content
- If No (0) to all TC and EIC indicators in Steps 1 and 2 ➔ The content is *lawful yet harmful*

Figure 3: Codebook visualisation



7.3 Lessons Learned from Scoring

With the OSINT research, 320 posts were collected and scored according to the pilot codebook elaborated in the previous paragraph. It is important to note that in the case of Reddit, 'posts' mean something different than on the other two platforms. On TikTok and Instagram, the OSINT research is limited to actual posts (consisting of a picture/video and accompanying caption), for Reddit, the OSINT research access is broader so 'posts' are threads (consisting of a caption and often accompanied by a picture and/or an external link) in subreddits, as well as comments on such threads.

Below, we will first provide an overview of the results of the coding. Next, we will reflect on the reliability of the coding and hence the detectability and identifiability of online content as terrorist, illegal, or implicit extremist content. We will also illustrate the challenges encountered in scoring some of the posts, in order to answer the second research question, namely: In what way, and to what extent, are the characteristics of extremist and terrorist content online, either or not in combination, detectable and identifiable in the online content?

7.3.1 Results OSINT Data Coding

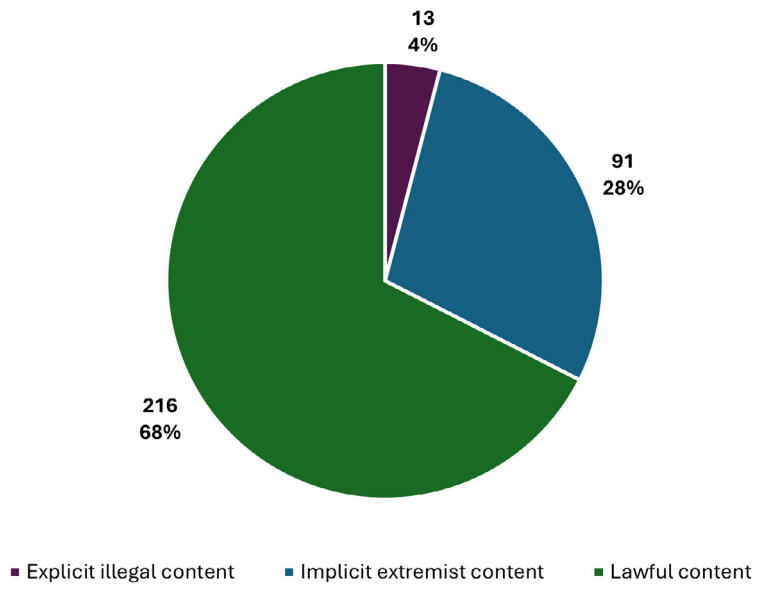
The coding of the 320 posts collected during the OSINT research phase resulted in the following scores:

Table 6: Qualification and indicators scores

Qualification/indicators	Indicator score	Qualification score
Presence of terrorist content (TC)		**0**
Presence of explicit illegal content (EIC)		**13**
Presence of hate speech (HS)	*9*	
Presence of incitement to violence (IV)	*1*	
Presence of denial, downplaying or justification of international crimes (DIC)	*0*	
Presence of both hate speech and incitement to violence	*3*	
Presence of implicit extremist content (IHC)		
Option 1 (1 indicator)	*195*	
Option 2 (at least 2 indicators)	*122*	
Option 3 (Concealing meaning + at least 1 indicator)	*61*	
Option 4 (Intent to cause harm + at least 1 indicator)	*91*	

No terrorist content was identified in the sample of OSINT data. This is not surprising, since hosting service providers (HSPs) in general will remove this kind of content in line with the TCO (either because it is in violation of the TCO and also covered by their Terms of Service (ToS), or due to an issued removal order). Almost two thirds of the content (68%) was not considered to be problematic or harmful at all and thus falls under the freedom of speech.

Figure 4: Distribution of types of content as described by the codebook found in the sampled posts from TikTok, Instagram and Reddit in percentages (n(t)=320)



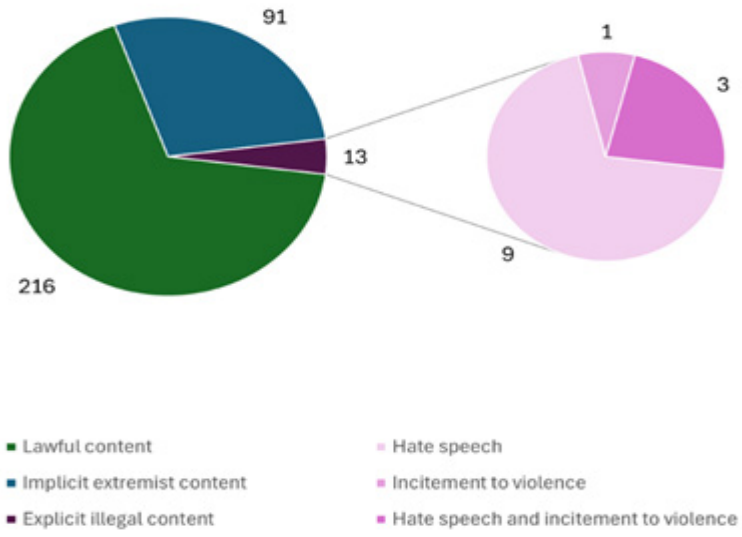
Keeping in mind the criticism related to the guidance provided by the legal definitions applicable to the category of explicit illegal content, it is not surprising that – despite the fact that according to the DSA the HSPs have to remove this kind of content - the research team still found 13 posts among the sample of 320 that were qualified as hate speech and/or incitement to violence.

Some of these posts can undoubtedly be qualified as hate speech. For instance, post #286, in which the headline of the post stated 'Pro-Palestine rioters destroy market booths and throw fireworks to invisibly provocative Maccabi supporters', the user comments: "Send all Arab

scumbags to Gaza. So they swallow bullets there". This post is thus a false negative.

Figure 5 - Post #286

Figure 6: Distribution of types of content as described by the codebook found in the sampled posts from TikTok, Instagram, and Reddit, zoomed into the distribution of types of explicit illegal content (n(t)=320)



Although some of these posts ticked the boxes relevant to the qualification of hate speech and/or incitement, it is nevertheless understandable that some of these posts were not removed. For instance, post #205 in a comment on the cited headline "Netanyahu spreekt over 'zeer gewelddadig incident' tegen Israeliërs in Amsterdam" (Translation: Netanyahu comments on 'extreme violence' against Israelis in Amsterdam), the user expresses his/her regret that the police is not shooting with sharp against the "Muslims scum". Following the indicators for hate speech and incitement to violence, the research team qualified this post to fall in both categories. The indirect phrasing by expressing a sentiment or regret might, however, have caused the 'false negative'.

Figure 7: Post #205

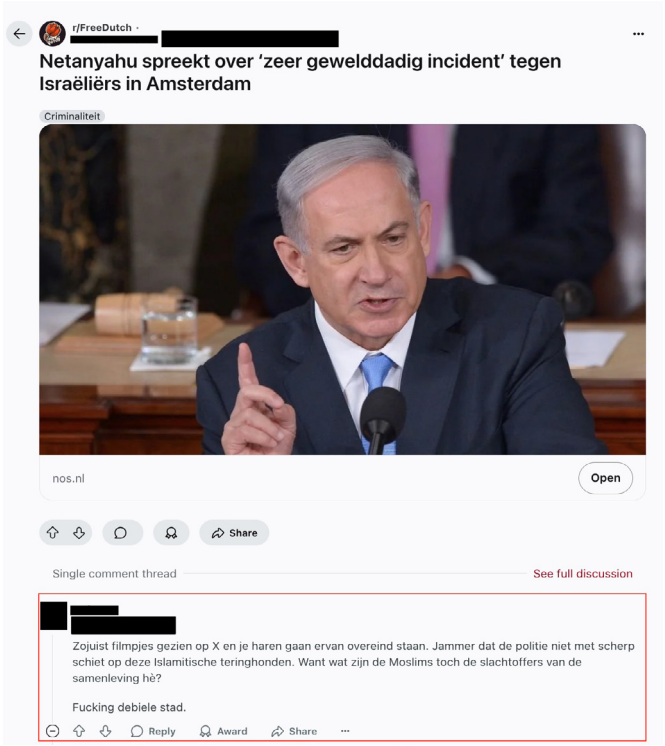
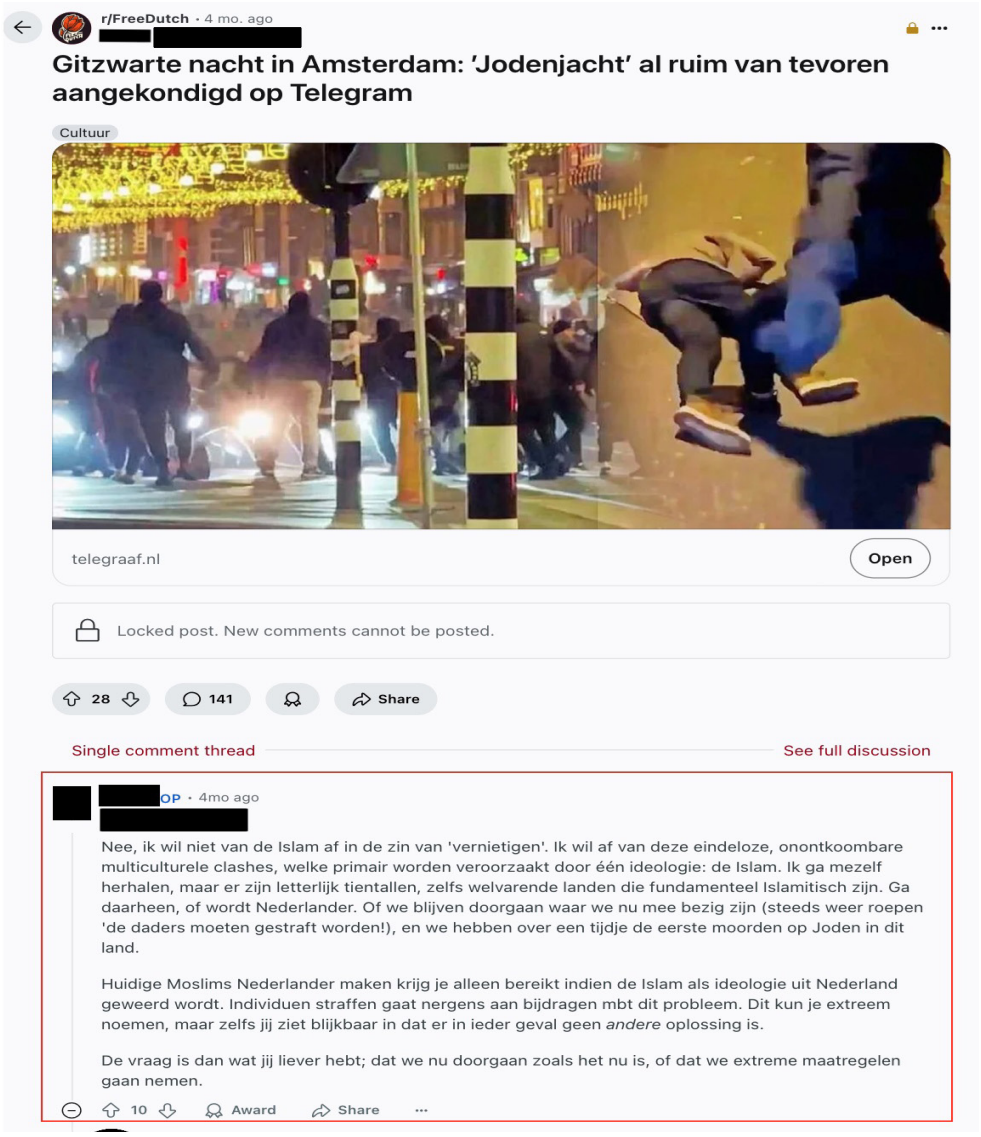


Figure 8: Post #223

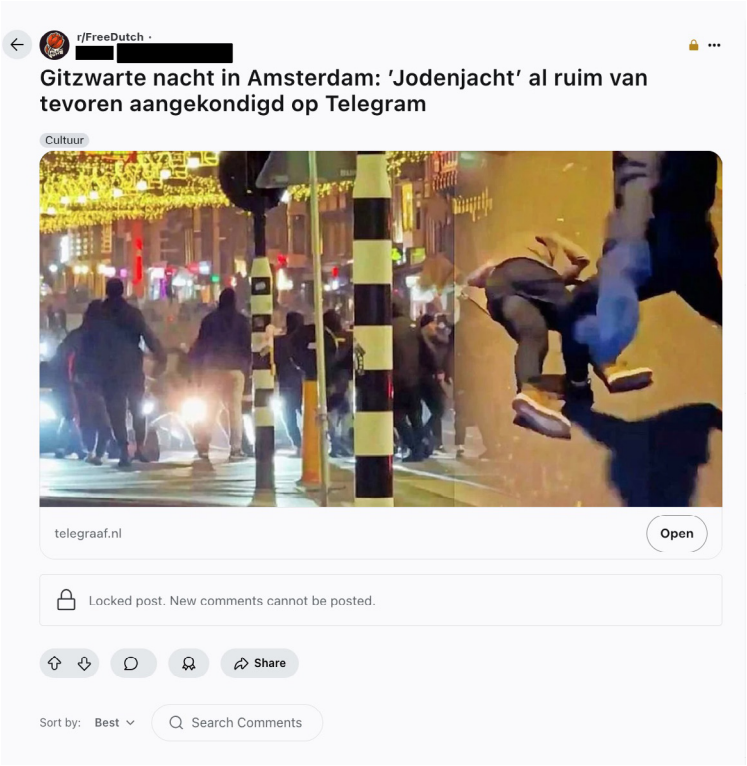


In another post (post #223), qualified by the research team as hate speech, the same user is debating the (in his/her mind) best policy to solve 'the issue with the Muslims in the Netherlands'. The user seems to know his/her point of view might be considered extreme and also claims not to be against Muslims or Islam in general. He/she, however, opines that there is no place for Muslims in the Netherlands. One might, however, argue that this post should still fall under the freedom of expression, and therefore was rightfully qualified by the HSP as a 'negative'.

Figure 9 - Post #245

Particularly, posts that frame the posts as if being part of a mainstream political debate seem to be a reason not to flag them as 'positive', and thus remove them for falling under the category of hate speech. An example is post #245. In the post, which asks users to give their opinion of a mostly left-wing political party with a strong agenda supporting immigrants in the Netherlands, a comment placed suggests that the fact that this political party is part of the political spectrum is exactly the biggest problem, due to their policy of allowing mass immigration from Muslim countries. The research team scored this post as falling under hate speech.

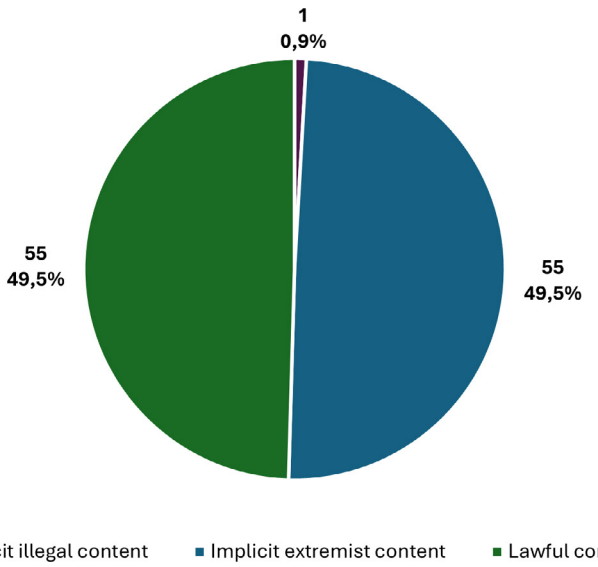
Figure 10 - Post #216



Particularly when assessing posts on Reddit, the Research team especially marked comments on these threads as illegal or implicitly extremist.[432] In some situations, it is actually the user who first posts a thread that is lawful (see post #216), but follows up with his own comment on this thread, which is subsequently marked as illegal (see post #223 in Figure 8 above).

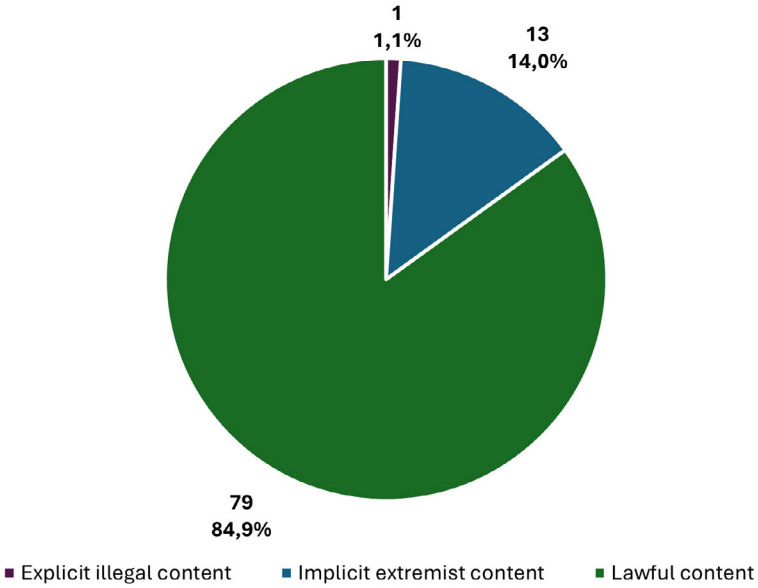
Comparing the scores of the OSINT data across platforms, it occurs that, particularly the data retrieved from TikTok is scored for almost half of the data to qualify as implicit extremist content. It is important to state that 97 percent of the posts fell in the category of far-right ideology.

Figure 11: Distribution of types of content as described by the codebook found in the sampled posts from TikTok in percentages (n(t)=111)



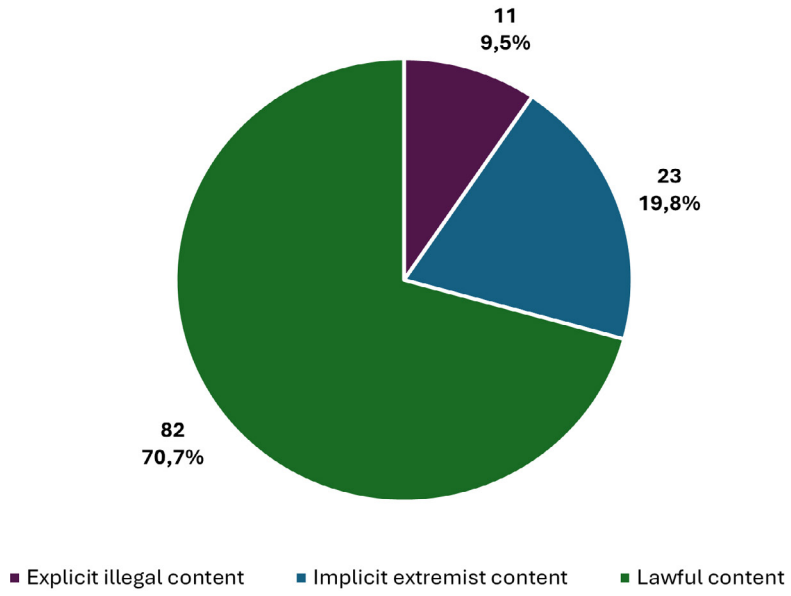
The presence of implicit extremist content was far less within the OSINT data retrieved from Instagram (14 percent), or Reddit (19,8 percent). The ideological qualification of the data for both platforms was also more diverse. For Instagram, 48 percent of the data referenced unclear ideologies, 40% to far-right ideologies, and 12 percent to Islamist ideologies. For Reddit, 80 perxent of the data reference to far-right ideologies, and 20 percent to Islamist ideologies.

Figure 12: Distribution of types of content as described by the codebook found in the sampled posts from Instagram in percentages (n(t)=93)



432 Reddit posts consist of threads and comments in threads. Of the 23 Reddit posts flagged as implicit extremist content, 2 were found in threads, and 21 in the comments in the threads. Of these 21 comments, 3 were made by the original user himself.

Figure 13: Distribution of types of content as described by the codebook found in the sampled posts from Reddit in percentages (n(t)=116)



Breaking down the qualification of the OSINT data per ideology, it is especially telling that the percentage of far-right posts qualified as implicit extremist content is significantly higher (33.6 percent) than this qualification among Islamist content (11,8 percent), or content with an unclear ideological affiliation (14,6 percent). This suggests that either the automated or manual detection and moderation mechanisms in place are better 'trained' to detect Islamist implicit extremist content and -to some extent- unclear ideologically affiliated implicit extremist content, or this has to do with users' behaviour. Islamist extremists might be more conscious about what they place on these three platforms, or extreme right-wing users refer more to attempted concealing methods.

Figure 14: Distribution of types of content as described by the codebook found in the sampled posts from accounts classified as 'far-right-inspired' in percentages (n(t)=238)

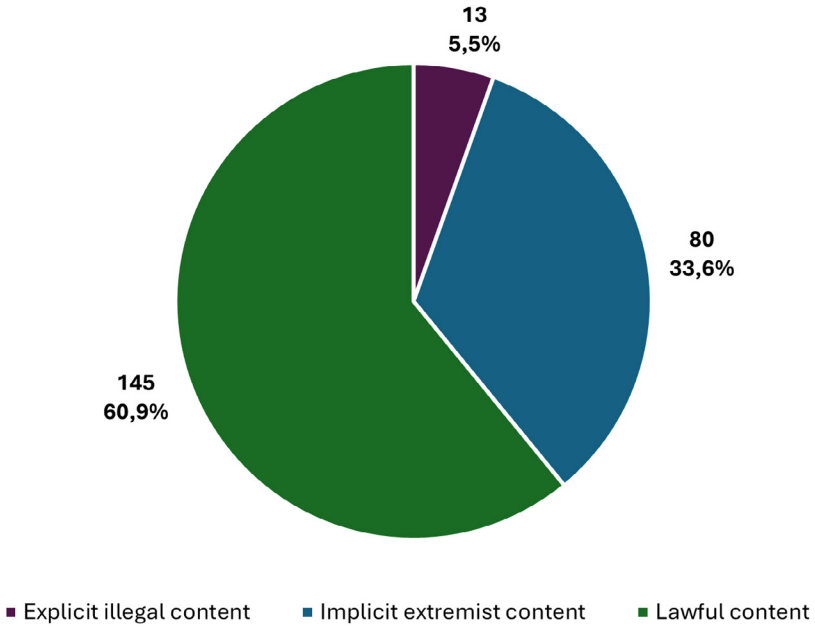


Figure 15: Distribution of types of content as described by the codebook found in the sampled posts from accounts classified as 'Islamist-inspired' in percentages (n(t)=34)

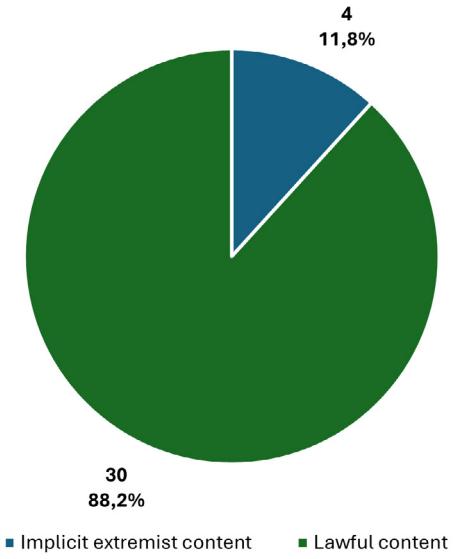
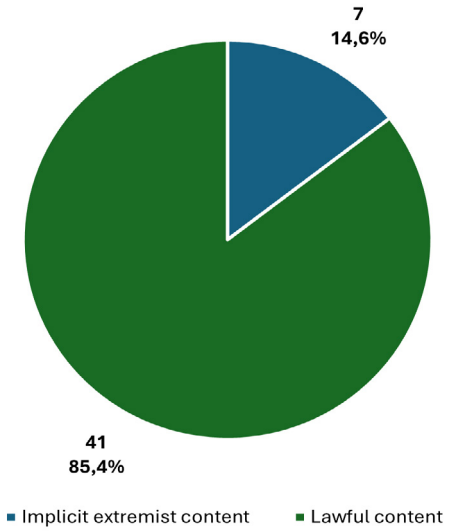


Figure 16: Distribution of types of content as described by the codebook found in the sampled posts from accounts classified as 'unclear ideological affiliation' in percentages (n(t)=48)



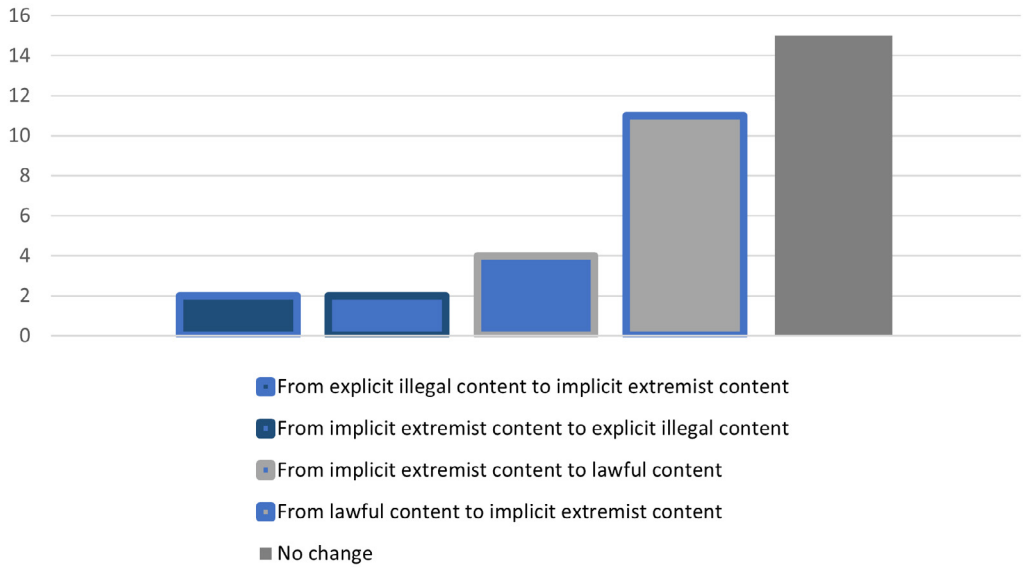
7.3.2 Challenges and Reliability of the Scoring

To test the reliability of the scoring, approximately 10 percent of the posts have been used for intercoding, meaning that another member of the research team would code a post that was previously coded by the researcher doing all the coding. Due to budget constraint it was not possible to intercode a larger sample of posts, nor was it possible to use a six eyes principle. The sample of 34 posts though was representative of the different languages (English, Dutch) and platforms (TikTok, Instagram, Reddit) used. Though representative of the encountered languages and selected platforms, this sample was not randomly selected. Since the main purpose was to test consistency and expose biases and blind spots, 26 posts of the sample were selected with this in mind and thus focused on posts that were deemed hard to code or highly interpretative.

The intercoding resulted in some changes in the scoring of the posts. In the figure (Figure 17) below, the number of changes is represented. Four posts either moved from the qualification from explicit illegal content to implicit extremist content (2), or the other way around (2). Considering

the fact that once implicit extremist content is qualified, in accordance with the pilot codebook, the posts need to be re-assessed to score whether they would nevertheless fall under terrorist or illegal content, these posts only show potential ambiguity related to the level of concealed language. Four posts, however, were initially qualified as implicit extremist content, yet after intercoding, nevertheless marked as falling within the freedom of expression. A total of 11 posts were, after intercoding, qualified as implicit extremist content instead of the initial score of lawful content. These posts can, of course, after following the feedback loop for the second assessment regarding the presence of terrorist or illegal content, once again be qualified as containing lawful content. Below, we will elaborate on the ambiguities that can be behind these differences in coding.

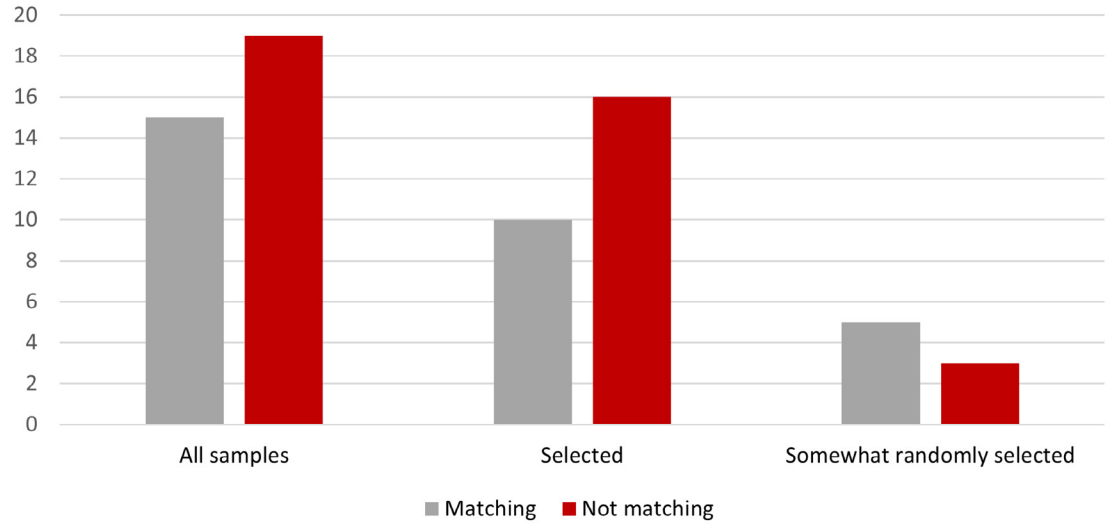
Figure 17: Changes in final evaluation of content between coders in the intercoding process (n(t)=34)



Apart from equipping the main coder with more expert guidance in the coding process, this intercoding process yielded some interesting results, providing us with a list of commonly mismatched indicators (i.e. indicators that did not correspond between coders for ten or more instances). This displays some technical and practical challenges to the feasibility of developing an assessment framework for implicit extremist content.

For instance, in a majority of samples, the coding result did not match between coders. This might hint at some practical and ethical challenges to the feasibility of developing an assessment framework for implicit extremist content. However, we must again stress that 26 of the 34 posts in the sample were especially selected for their interpretability. Of the eight cases that were randomly selected, a majority of cases did match.

Figure 18: Results of the intercoding process divided by type of selection (n(t)=34; n(s)=26; n(srs)=8)



The table below (Table 7) provides an overview of the scoring discrepancies between the main coder and the coding done by the other members of the research team. Per indicator used, a count is kept of the number of times the main coder marked that indicator with a [1] (indicating the indicator was present in the post). The table also keeps a record of the number of times of the maximum 34 intercoded posts, where there is a mismatch between the main coder and the coding by a research team member. Mismatches could occur either because there is a difference of interpretation, or because the members of the research team doing the intercoding missed crucial background information to make the right assessment. The latter could, for instance, occur if by coding a singular post of an account, the research team member missed the broader picture of the context of the post, which would have become clear if all posts of that account had been coded by the same person. The information on the mismatches, in addition to the coder's own experience with the coding, resulted in the colour coding per indicator regarding the usability, usefulness and interpretability of each indicator.

The usability relates to how easy it is to code the indicator, or how user-friendly it is. For instance, if common knowledge is enough to code the indicator, and it is not very time consuming to make the assessment, because no further background research is needed to make the assessment, the usability would be marked as green. When some more effort is needed, yet the needed background to make the assessment is easy to retrieve, the indicator would be marked orange. And when it is very difficult or time consuming to make the assessment, the indicator would be marked red for usability.

The usefulness of an indicator provides information on whether the assessment of the indicator is absolutely necessary to come to an overall good assessment of the category of terrorist, illegal or aspects of implicit extremist content. Red suggests the indicator is already sufficiently covered by another indicator, or does not contribute significantly to the overall assessment, and can be removed. Orange indicates the indicator is hardly ever used, or has some overlap with other indicators.

The interpretability of the indicator provides insight into the level of subjectivity in the interpretation of the indicator. With more than 10 mismatches (in particular, the ones related to a difference of interpretation as opposed to the mismatches due to lack of broader contextual understanding) among the intercoded sample, the indicator would be marked as red, meaning that the indicator is considered highly subjective. When there are 1-10 mismatches, the colour code orange indicates that the formulation of the indicator needs to be improved to avoid subjectivity in interpretation.

When marked as green, no problems with the interpretability of the indicator were discovered. In the table (Table 7) below, the indicators are grouped per subsection of terrorist or illegal content or elements of implicit extremist content, except for the first seven, which are used for different categories, and which are marked in the table as crosscutting indicators, or which are similar in meaning. Where relevant, more elaborate explanations regarding the assessment of the usability, usefulness and interpretability of the indicators are offered below.

Table 7: Usability, usefulness, and interpretability scores for indicators

INDICATORS	USABILITY	USEFULNESS	INTERPRETABILITY	COUNT	MISMATCHES (34)
Crosscutting and similar indicators	-	-	-		
Intended audience is a **group** (IT/IV)				178	19 (1/2)
Content is spread in **public** (IT/HS/IV/DIC)				320	0
Awareness of likelihood that the audience could commit a terrorist act / serious violent offence - **capacity** (IT/IV)				293/296	31
Knowledge of or reference to terrorist activities (IT/GT)				10	6
Not a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (IT/GT/RT/HS/IC/DIC)				242	10
Audience sufficiently clear to include people with **ideological ties** (IT)				239	26
Directly or indirectly addressing an audience - individual or group (RT)				319	19 (2/2)
Incitement to commit/participate in a terrorist offence (IT)	-	-	-	**0**	**0**
Advocate the commission of a terrorist offence (IT)				0	0
Deliberately creates the likelihood of a **serious** terrorist offence regardless of whether executed or not (IT)				6	0
Direct or indirect **causal link** between incitement and terrorist act (IT)				0	0
Imminent risk that a terrorist act will be committed (IT)				0	0
Glorification of a terrorist act (GT)	-	-	-	**0**	**0**
Justification of past/future serious terrorist offence (GT)				2	0

Deliberately creating an **atmosphere of fear** (GT)				65	11
Recruitment for a terrorist organisation (RT)	-	-	-	**0**	**0**
Call to supply information/ material resources/funding/human resources to a terrorist group (RT)				0	0
Knowledge of or reference to a specific terrorist group, one or more serious criminal acts committed by the referenced terrorist group, one or more concrete aims of the terrorist group (RT)				2	1
Hate speech (HS)				**12**	**4**
Discriminates a group or individual (HS)				37	8
Discriminates **on the basis of** race, gender, religion, belief, sexual orientation, physical or mental disability (HS)				20	4
Insults/incites hatred or violence (HS)				116	13
Knowledge of or reference to discriminatory acts or speech (HS)				26	11
Incitement to violence (IV)				**approx. 4**	**1**
Direct and imminent causal link between incitement and serious violent offence (IV)				7	1
Advocates commission of a serious violent offence (IV)				8	0
Deliberately creates the likelihood that a **serious** violent offence will be committed regardless of execution or not (IV)				54	10
Knowledge or reference to serious acts of violence (IV)				approx. 97	9
Denial, downplaying or justification of international crimes (DIC)				**0**	**0**
Against a **group or individual** (DIC)				184	10
Denies, justifies, or downplays core international crimes established irrevocably by a court (DIC)				2	0
Knowledge or reference to the fact that crimes constitute core international crimes (DIC)				9	0
Intent to conceal meaning (CM)				**77**	**8**
Ambiguous or coded language (CM)				25	4

Altered text or misspelling (CM)				7	0
Misleading cover image for video (CM)				1	4
Blurred or altered image - e.g. marker function to hide certain words (CM)				7	0
Use of humour or irony (CM)				39	6
Harmful alliances/affiliations (HA)				**31**	**7**
Hateful / extremist symbols (RWX)				3	0
Hateful / extremist symbols (Islamist)				0	1
Hateful / extremist emojis (RWX)				0	0
Hateful / extremist emojis (Islamist)				0	0
Hateful / extremist coded slogans, slang, or acronyms (RWX)				13	1
Hateful / extremist coded slogans, slang, or acronyms (Islamist)				0	0
Problematic references to historical/current context (PR)				**86**	**14**
Denial or questioning of proven past or present crimes (PR)				9	2
False or misleading claims about past or present crimes (PR)				33	2
Justification of current or potential future crimes through references to past crimes (PR)				13	3
Falsified historical claims aimed at denying the existence or territorial legitimacy of a state (PR)				5	2
Glorification or positive portrayal of individuals or groups involved in crimes (PR)				24	3
Glorification or positive portrayal of public figures known for spreading hateful or extremist narratives (PR)				25	2
Promotion or endorsement of books, films, essays, or any other products known for spreading or supporting any of the above narratives (PR)				24	3
Implicit action triggers (AT)				**125**	**10**
Instilling a perceived need for retaliation (AT)				23	6
Instilling a perceived need for self-defence (AT)				96	10
Instilling a perceived need to protect a group from reputational damage (AT)				33	3
Instilling the idea of compensation or reward for undertaking action (AT)				2	-

Exerting peer pressure/respect for a code of honour (AT)				18	4
Providing sacred justification for action (AT)				11	3
Presumed intent to cause harm (IH)				**107**	**12**
Normalise hateful/violent narratives (IH)				71	10
Propagate hateful/violent conspiracies (IH)				46	11
Foster hate/hostility towards an out-group (IH)				37	17
Trigger to take online harmful actions - e.g. harassment/doxxing (IH)				2	0
Trigger to take offline harmful actions (IH)				32	4

Challenges with *crosscutting and similar indicators*

A couple of indicators are used for different categories of terrorist, illegal, or implicit extremist content. Reflecting on the usability, usefulness and interpretability of these indicators, we identified several challenges with some of the indicators.

Intended audience is a group: The term 'group' causes confusion. The main coder initially coded all messages that could conceivably be intended to reach an audience of more than one - all public messages - as valid. After intercoding, and noticing the number of mismatches, the main coder switched to only counting messages referring to an (imagined) group and messages posted in a group (sub-reddit) as valid. That, however, triggers other challenges. There are explicit and more implicit ("we", "our people/country", through hashtags) ways to refer to groups and it is hard to determine when a message within a sub-reddit is truly meant only as a response to another user or if the format of a reply is used to reach a broader public - and the intended audience is thus a group.

Awareness of the likelihood that the audience could commit a terrorist act/serious violent offence – related to perceived capability: Whether an audience has the capacity to commit a terrorist or violent act is highly interpretative, and answering the question requires a lot of contextual knowledge about the user, their followers, the group they post in and their other content. The main coder's bar for capacity was relatively low, since terrorist and violent acts are quite regularly conducted by actors with little capacity and know-how. Every post from a user with terrorist, or implicit or explicit extremist connotations, a profile description and/or picture which suggests ideological commitments, or a targeted audience of politically conscious youth and/or adults, validated this indicator.

Not satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation: This indicator is quite subjective and takes a lot of time to code. Since anything can really be stated in the form of a 'satire, parody, artistic expression, it is tough to determine whether or not something actually is any of those things. A racist caricature is not any of those things per se, but it can be all of those. Whilst no (legitimate) commemoration of historical events, colonial past or decolonialisation was present in this sample, the main coder believed it would be quite difficult to determine whether

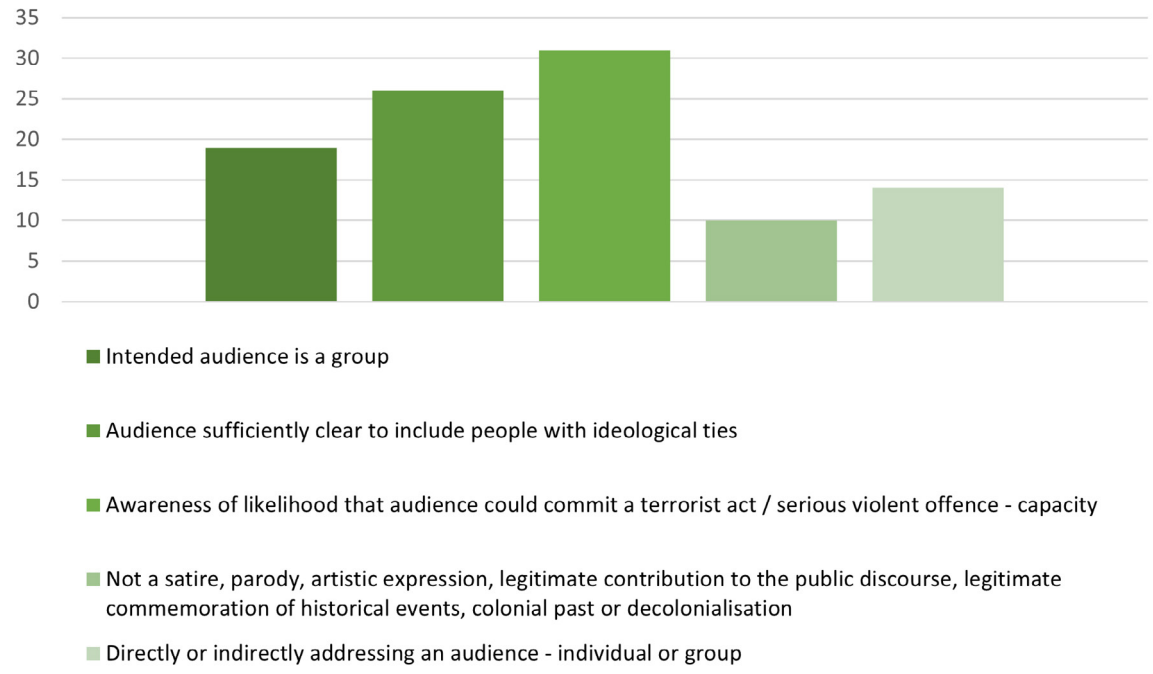
or not such a commemoration is legitimate. Legitimate contribution to the public discourse is similarly hard and subject to opinion. The main coder chose to only include contributions that were factual, however since sitting politicians and mainstream news platforms quoting them, nowadays quite often broadcast views containing counterfactuals, this indicator remains challenging.

Audience sufficiently clear to include people with ideological ties: This indicator requires a lot of background knowledge on the ideology of the users and their followers/viewers/groups. Whilst it is not that hard to determine whether the audience includes people that are ideologically tied to the subject, this requires a bit of knowledge and research into the user and their audience - some groups or hashtags are very clearly ideologically centred, whilst others are more general but still include the intended ideological audience. In this coding, users with an unclear ideology would automatically score a [0] on this indicator. Even though a non-Islamist, non-left wing extremist pro-Palestine activist that would call for the assassination of Zionists in government, should/would be considered as validating as incitement to terrorism. To mitigate this, we might need to clarify what ideology means in this exact context. Furthermore, the trend toward non-ideological extremism also implies that there is no clearly delineated audience that has clear ties with a specific ideology. Instead, a mere fascination with violence might suffice.

Directly or indirectly addressing an audience – individual or group: The main coder interpreted this indicator as a more specific formulation of the indicator 'Content is spread in public', since all content that is spread publicly is at least indirectly addressed to someone, unless the content is spread publicly but only really meant for personal use. Another case of unique usefulness would be if some samples included non-publicly spread posts that did address an audience - like personal messages (in private groups) or posts by private accounts. Since this is not the case for our sample, this indicator is not that useful or might require refinement.

The figure (Figure 19) below provides insight into the number of mismatches per crosscutting or similar indicator.

Figure 19: Most common mismatched crosscutting indicators in the intercoding process (n(t)=100)



Challenges with scoring indicators for terrorist and illegal content

Reflecting on the usability, usefulness and interpretability of the indicators used to score whether content could be qualified as terrorist or illegal content, we identified some challenges with some of the indicators.

Indicator related to incitement to commit/participate in a terrorist offence

Deliberately creates the likelihood of a serious terrorist offence regardless of whether it was executed or not: It is considered a bit hard to determine whether or not users deliberately create a likelihood or whether their anger, conviction or bigotry blinds them to this. Additionally, whilst it is quite possible to determine - although still subject to opinion - whether or not a message creates fear or moves people to commit a violent offence, weighing whether or not a statement is strong enough to create the likelihood of a serious terrorist offence being committed is much harder and probably requires more expert knowledge.

Indicator related to glorification of terrorist offences

Deliberately creating an atmosphere of fear: This indicator is highly interpretative and dependent on personal biases. However, since terrorist content and explicit illegal content indicators require validation of all sub-indicators, the fact that this indicator is relatively subjective is not necessarily problematic.

Figure 20 - Post #024

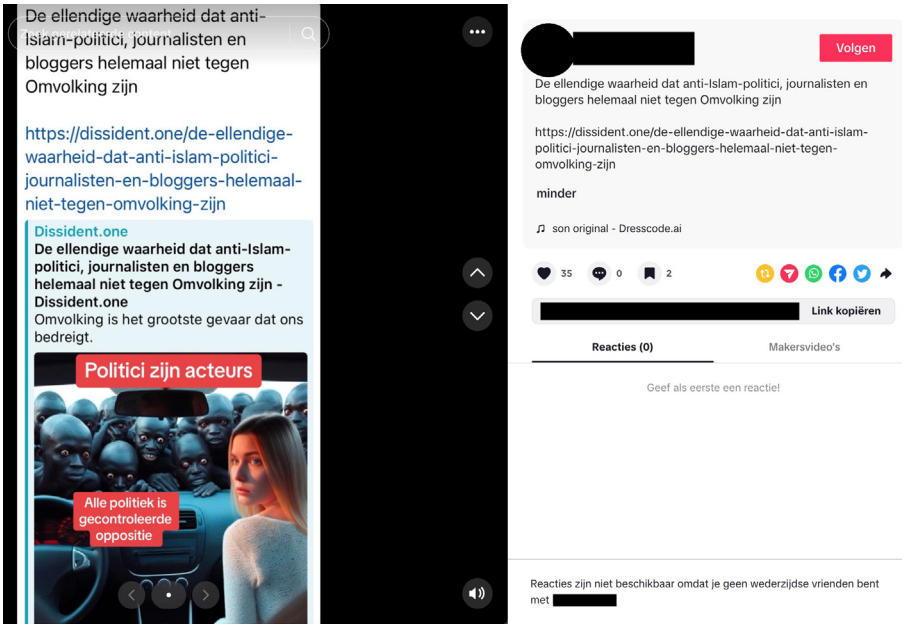
Indicators related to hate speech

Discriminates a group or individual: What constitutes discrimination is hard to determine, especially considering that not everyone may be expected to instantly notice or understand racist dog whistles and caricatures. Without a closer look and contextual clues Post #024 is just an anti-capitalist message. Alternatively, whilst experts might note that Post #031 is a clear case of discrimination, and even a very specific racist stereotype, often used by white nationalists to instil fear, where black people are pictured stealing or badgering white women to give power to the great replacement theory, a judge might note that an AI-generated picture of black people looking at a white woman is not tantamount to discrimination.

Discriminates on the basis of race, gender, religion, belief, sexual orientation, physical or mental disability: This indicator could quite easily be combined with the previous indicator.

Insults/incites hatred or violence: Whether or not something is insulting or inciting hatred/violence is very subjective. It is not hard to validate this indicator when mainstream insults or slurs are used, or when a post explicitly incites hatred/violence, but in most cases - as is common with implicit extremist content - the insulting nature of the posts is somewhat concealed. With partial concealment, this indicator becomes highly interpretative, because one might 1) not find something insulting correctly - 0, 2) not find something personally insulting which is actually insulting - 0, 3) find something insulting - 1, 4) find something insulting but too concealed - 0, or find something insulting that is actually only insulting after one lifts the veil - 1.

Figure 21 - Post #031



Knowledge of or reference to discriminatory acts or speech: This indicator might not be very useful. References to discriminatory acts of speech are clear, but knowledge of them is almost impossible to determine. Whereas it can be argued that most people who commit discriminatory acts of speech do this knowingly, but many of them will claim their speech acts are not discriminatory to start off with, often citing tangential factors like culture or religion.

Indicators related to incitement to violence

Direct and imminent causal link between incitement and serious violent offence: Whilst this indicator seems adequately defined, some examples of direct and imminent causal links might be helpful to get a better understanding of which edge cases would fall within this indicator.

Deliberately creates the likelihood that a serious violent offence will be committed regardless of execution or not: This indicator is very interpretative, because whether a polarising political commentary deliberately creates a likelihood for violence is probably much dependent on the political positions one holds and one conceives of as conventional. Additionally, it is hard to determine whether or not users deliberately create a likelihood or whether their anger, conviction or bigotry blinds them to this.

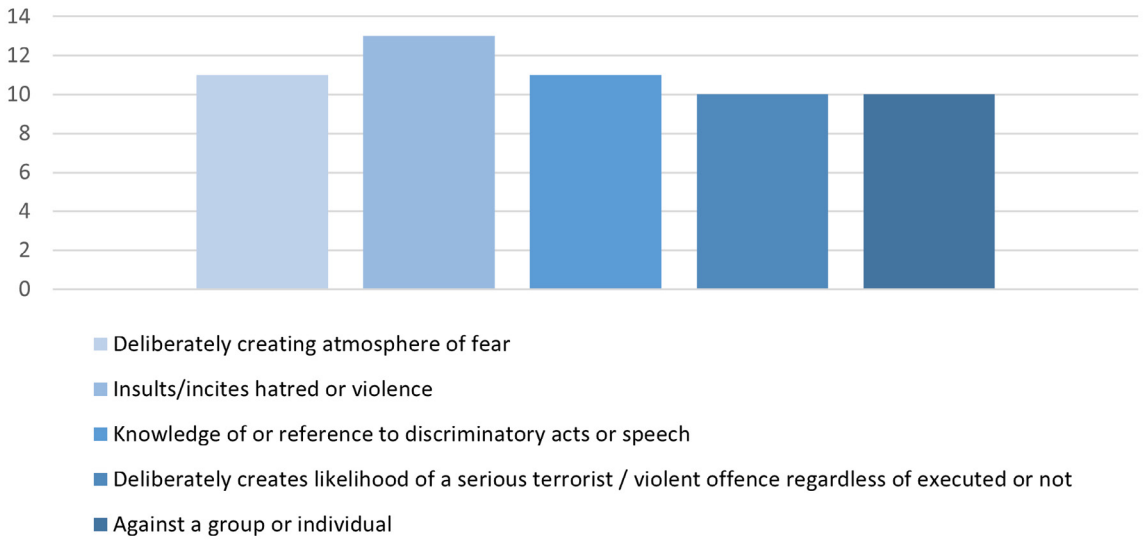
Indicators related to denial, downplaying or justification of international crimes

Against a group or individual: This indicator seems not very useful, as most borderline and extremist posts can be described as 'against' someone or something. It would probably be wise to clarify what 'against' means here or leave it out wholly.

Knowledge or reference to the fact that crimes constitute core international crimes: It is a little interpretative whether or not someone knows the referenced crimes constitute core international crimes.

The figure (Figure 22) below provides insight into the number of mismatches per indicator used to score terrorist or explicit illegal content.

Figure 22: Most common mismatched explicit illegal content indicators in the intercoding process (n(t)=45



Challenges with scoring indicators for implicit extremist content

Reflecting on the usability, usefulness and interpretability of the indicators used to score whether content could be qualified as implicit extremist content, we identified some challenges with some of the indicators.

Indicators related to concealing meaning

Ambiguous or coded language: Ambiguity is subjective and dependent on content knowledge. However, considering the fact that many indicators contribute to the overall assessment, the potential subjectivity of this indicator gets mitigated by the other indicators. Furthermore, a clearer definition of what entails ambiguity might further alleviate any potential problems.

Altered text or misspelling: The research team found relatively little use of altered text or misspelling in the sample data. Those posts that did, however, did not have the goal to conceal/

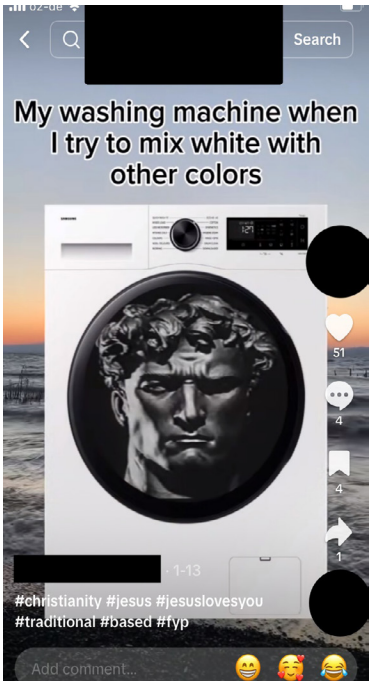
mask extremist rhetoric, but instead were not 'limited' by algorithms for using controversial words like 'arrestee'. This indicator could be expanded to include altered imagery with the express purpose of concealing meaning and evading filters, which would make it more useful.

Use of humour or irony: Humour is not always used to conceal a message, sometimes humour is simply humour and sometimes humour is very explicit in representing an extremist view and thus more akin to a harmful alliance.

Indicators related to harmful alliances/affiliations

In general, the use of coded symbols, emojis, and slogans/slang/acronyms was very limited in this sample. Many posts were either very implicit, seeking to use arguments that sound reasonable to further extremists' ends, or very explicit in their extremism, albeit using humour or neutral language to dress up their views. Relevant to note is that Dutch extreme right-wing users that we encountered in our sample did not often use the known English slogans/slang/acronyms. In fact, whilst they do follow more global trends of rhetoric (COVID-conspiracy, Great Replacement Theory, anti-establishment hate targeted towards liberal and progressive people and ideas), their language - especially on text-focused platforms like Reddit - seems somewhat unique to the Netherlands and removed from that of the English-speaking far-right. They often use Dutch contemporary sayings, memes, irony-posting and generally are quite explicit in their rhetoric - often calling for the deportation of Muslims and the ban on the religion, something that may have to do with the normalisation of this position by a radical right-wing political party. It is considered advisable – as part of a follow-up research- to look into the exact phrases, dog whistles and other extremist rhetoric that are used by the Dutch-speaking extremist groups. Meanwhile, in general, databases that keep track of known symbols, emojis, and slogans/slang/acronyms would constantly need to be updated. While a 'Roman salute' is, for instance, generally considered to be a reference to the Nazi salute, a picture of a head of a Roman statue might be less obvious as a symbol of right-wing extremism (post #003).

Figure 23: Post #003



Hateful extremist coded slogans, slang, or acronyms (of an Islamist nature): Whilst Islamist-inspired or adjacent users in this sample use ambiguous and coded language, this is almost never in the form of slogans/slang or acronyms. It seems like Islamist (-inspired) extremists are either more aware of the scrutiny they are under or believe 'grounded' analysis - often combining liberal

and progressive critiques with religious and cultural doctrine - is more effective. Alternatively, it could also be that one needs to be extremely well versed to understand the coded language they use. Furthermore, there are very little up to date databases, and whilst it was possible to translate Arabic text, that does not necessarily help in understanding the deeper meaning and origin of certain phrases and ideas. It may also simply be a sign of the time. It is our observation that far-right inspired extremists seem to feel inviolable and emboldened by a global surge of far-right governments. As a result, they become more forthcoming in their rhetoric. In contrast, Islamist-inspired extremists have been scrutinised for years by many governments, explaining the fact that our sample hardly included any explicit extremist content with an Islamist ideological background. However, given the fact that supporters of this ideology do currently have some social capital for being on the more popular side of some global conflicts, like the Israel-Gaza war, the posts found in our sample seem to suggest that there might be another strategy that is being followed: they simply attach themselves to the popular side of an international conflict and interweave their message with that of a popular anti-war movement.

Indicators related to problematic references to historical/current context

Justification of current or potential future crimes through references to past crimes: The term 'justification' might require some additional information, including examples for an improved codebook. While justification in references to past crimes might be less contested, whether references to the underlying causes of current crimes by referencing prior crimes committed might be qualified as an illegal justification of these international crimes, is less straightforward. Certainly not if even mainstream news media sometimes adopt this framing.

Falsified historical claims aimed at denying the existence or territorial legitimacy of a state: This indicator should be expanded to also include contemporary claims.

Glorification or positive portrayal of public figures known for spreading hateful or extremist narratives: This indicator is highly useful but does have interpretative limitations. Firstly, who is considered large enough to be a public figure is debatable. For instance, in some extremists' niches, having 10,000 followers would mean that one ticks that box. Secondly, whilst this indicator is perfect to encapsulate online extremist 'thought-leaders', the fact that some radical thought-leaders have established themselves in the heart of the political power in countries, makes this potentially a (politically) sensitive indicator to work with.

Indicators relevant for implicit action triggers

In general, these indicators are useful and contribute to making an overall assessment of whether the content qualifies as implicit extremist content. Yet, many are quite interpretative. Further clarification of the indicators is needed for more reliability in use.

Instilling a perceived need for retaliation: Some examples of what this looks like might be beneficial to get a more intersubjective understanding of 'retaliation'.

Instilling a perceived need for self-defence: Highly interpretative and dependent on personal biases. This is problematic because it easily opens for abuse by biased flaggers or coders who have an agenda to report this as problematic. There should probably be a range of examples of acceptable ways of describing threats to distinguish between those and posts that rightly fall under this category.

Instilling the idea of compensation or reward for undertaking action: This indicator has not appeared many times in the coded sample. The indicator might also overlap with the indicator on peer pressure and sacred justification

Exerting peer pressure/respect for a code of honour: This indicator could use some further clarification to improve its interpretability and usability. Many posts include some form of a call to action, but not usually through a code of honour or by exerting peer pressure. More commonly used are posts instilling fear or calling for justice. Furthermore, peer pressure does not always make sense for online groups in which relationships are not very deep.

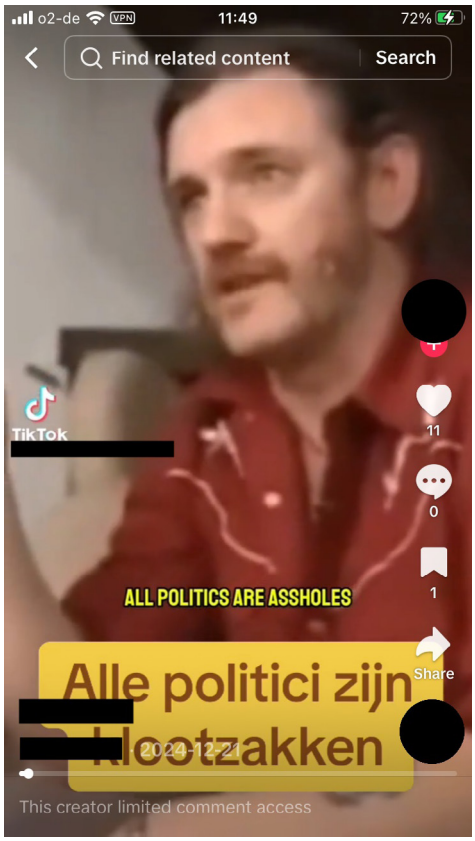
Providing sacred justification for action: While this indicator is sufficiently clear, it limits this indicator to Islamist-inspired posts. Using 'building on non-negotiable beliefs' might be more ideologically neutral, but also less clear.

Indicators related to presumed intent to cause harm

Due to the reference to 'presumed intent', these indicators are by default very interpretative. Some of that subjectivity is inherent to the concepts we are dealing with, and some of it could be decreased by further clarification of the codebook.

Normalise hateful/violent narratives: This is one of the most used indicators regarding the criteria of 'intent to cause harm'. It is very useful but inherently subjective and dependent on biases and content knowledge.

Figure 24: Post #016



Propagate hateful/violent conspiracies: There is so much overlap between the indicator of hateful narratives (falling under the previous indicator) and this one, focusing on conspiracies, that it would make sense to merge these indicators into one. Although this would not inherently make these indicators less interpretative, it would make them more usable and decrease mismatches between coders.

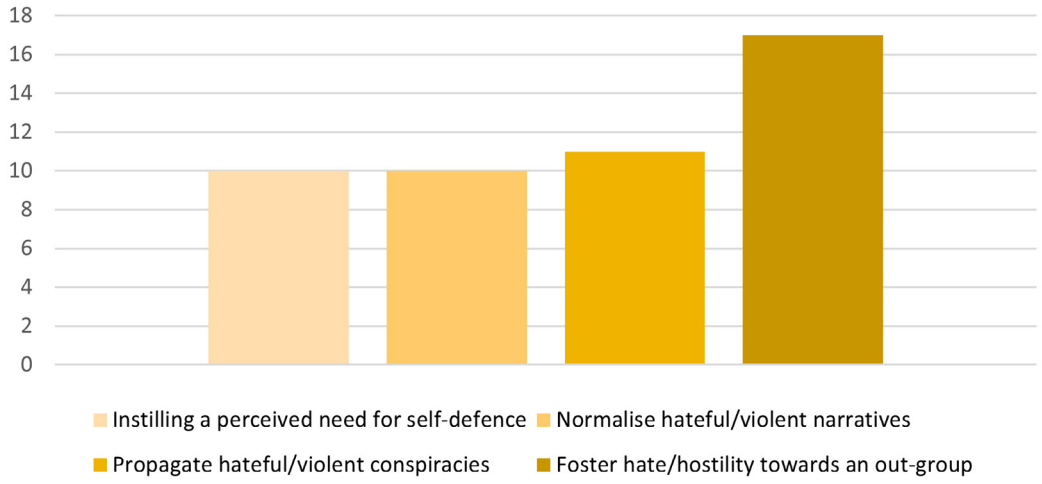
Foster hate/hostility towards an out-group: This is potentially a very useful indicator. However, there are two challenges that need to be addressed. Firstly, it would be necessary to define

what qualifies as an out-group. Some clarification and examples would tremendously ease the use of this indicator. For instance, far-right inspired users sometimes blame migrants for the problems in society and clearly frame them as an out-group. However, fostering hate targeted towards the government, the question is whether that would tick this box. See, for instance, the example of post #016. Secondly, the creation of in- and out-groups is pervasive in political communication, and even the fostering of some 'constructive' hate towards such out-groups is not necessarily intended to cause harm. Where to place the threshold, in other words, can sometimes be debatable.

Trigger to take offline harmful actions: What a harmful action entails requires some clarification.

The figure (Figure 25) below provides insight into the number of mismatches per indicator used to score implicit extremist content.

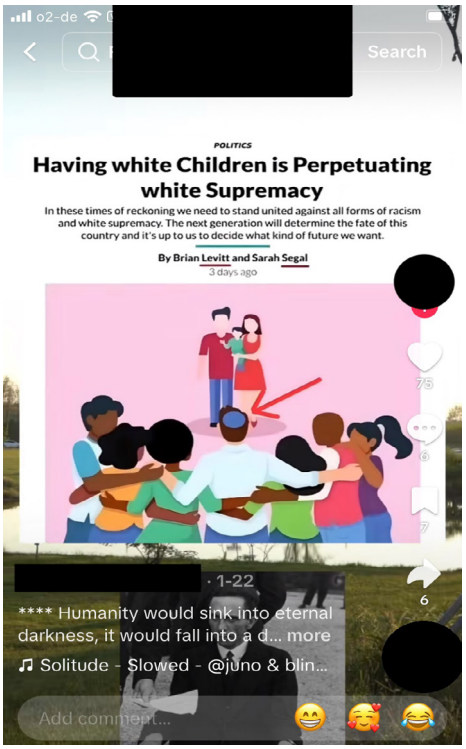
Figure 25: Most common mismatched implicit extremist content indicators in the intercoding process (n(t)=48)



7.3.3 Indicators Deemed Missing in the Pilot Codebook

Within the category of implicit extremist content, each group of indicators included an 'other' category. Based on the experiences in coding the data of the OSINT Research, extra indicators could be included to assist in the coding for that group of indicators.

Figure 26 – Post #007



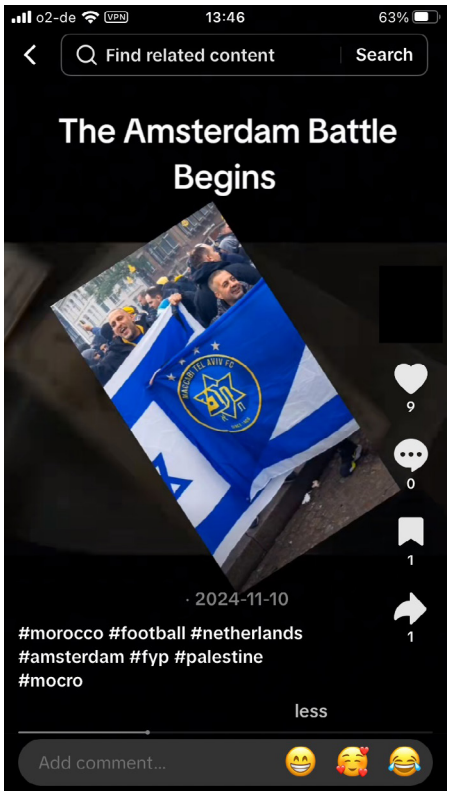
Concealing the meaning – indicators missing

The indicator assessing the existence of humour should also assess the existence of non-ideological 'inside' jokes. It would also be useful to include an indicator related to the visualisation or caricaturising of an out-group in order to avoid naming them in the text. See, for instance, post #024 in Figure 20 above. Related to this, it would be important to also look at altered imagery to obscure or highlight hateful content. An example of the latter can be found in post #007.

The use of hateful analogy as a way to conceal the meaning is currently also not sufficiently covered in the indicators. See, for instance, the use of a washing machine as an analogy for race-mixing, or a reference to Batman as an analogy for taking vigilante justice on hooligans (see posts #003 and #111 – Figure 23, which is included above, and figure 27 below).

Furthermore, oftentimes the meaning is concealed by merely referencing a link toa post, video, or other article with problematic content either on the same or on an external site or platform. And finally, asking a question as a tactic to conceal the hateful or extremist normative statement, is also commonly used.

Figure 27: Post #111



Harmful alliance or affiliation – missing indicators

To complement the indicators assessing the presence of hateful/extremist symbols, emojis, coded slogans, slang or acronyms, it would be useful to add: hateful/extremist coded pictures, racist caricatures and tropes of a certain group of people, or AI-generated pictures. Furthermore, this subset of indicators would benefit from adding indicators assessing hateful/extremist references to conspiracy theories or slurs.

It would, furthermore, be important to also assess the captions used and to reflect on the quotes included. In post #007 (included above in figure 26), for instance, the quote referenced is taken from “The Creators of the World’s Misfortunes” by Joseph Goebbels.

7.4 Observations

7.4.1 Necessary Combination of Characteristics of Terrorist and Illegal Content

To analyse which combination of characteristics is needed to determine whether content qualifies as terrorist content, we focused on the legal elements in the TCO relevant for incitement to commit or participate in terrorist offences, glorification of a terrorist act, and recruitment for a terrorist organisation. All of these indicators needed to be present to qualify the content as terrorist content. These indicators would need to be included in an assessment framework that would check the presence of these indicators before drawing conclusions on whether the content can be qualified as terrorist content, and would need to be moderated.

For a comprehensive framework to identify whether content qualifies as terrorist content, indicators should also be developed for the following acts:

- Threat to commit a terrorist offence as provided in Art. 3 (1) lit. a-I EU Directive 2017/541
- Instructions on the making and use of explosives, firearms or other weapons or noxious

or hazardous substances, or on other specific methods or techniques for the purpose of committing a terrorist offence
- Terrorist financing

To analyse which combination of characteristics is needed to determine whether content qualifies as illegal content, we focused on the elements of the Dutch criminal code for hate speech, incitement to violence, and denial, downplaying or justification of international crimes. All of these indicators needed to be present to qualify the content as illegal. These indicators would need to be included in an assessment framework that would check the presence of these indicators before drawing conclusions on whether the content can be qualified as illegal, and would need to be moderated.

For a comprehensive framework to identify whether content qualifies as illegal content, indicators should also be developed for the following acts:

- Defamation
- Child sexual abuse material
- Weapons offences

7.4.2 Detectability and Identifiability of Terrorist and Illegal Content

The indicators used were tested regarding their usability, usefulness, and interpretability. Of the 26 indicators used (some are used for multiple offences), ten are considered to be usable, useful, and easy to interpret. Eight indicators need some attention or adjustment to ensure an accurate use of the indicators.

More serious concerns arise regarding the following eight indicators:

- **Intended audience is a group**
  - *Usability:* The ambiguity of the term *group* makes this hard to apply consistently. The coding strategy had to shift mid-process due to high inter-coder disagreement.
  - *Usefulness:* The indicator is central to assessing the reach and intent of the message, but the usefulness is reduced when the concept of *group* is inconsistently applied or understood.
  - *Interpretability:* Highly interpretative - group references can be explicit ("we", "our people") or implicit (hashtags, subreddit format). Determining true intent (reply vs. group broadcast) is context-heavy and subjective.

- **Awareness of the likelihood that the audience could commit a terrorist act/serious violent offence – related to perceived capability**
  - *Usability:* Low - requires significant contextual understanding of user history, group dynamics, ideology, etc. The coder used a low threshold due to the unpredictable nature of such acts.
  - *Usefulness:* Somewhat useful but risky - broad application could lead to over-inclusion, while stricter definitions may lead to underreporting.
  - *Interpretability:* Very interpretative - dependent on ideological assumptions and perceptions of user demographics. Needs clear criteria and examples in the codebook.

- **Not satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation**
  - *Usability:* Very low - ambiguous, slow to apply, and heavily dependent on the coder's cultural and political background.
  - *Usefulness:* High in theory (helps avoid false positives), but low in practice due to how easy it is for harmful content to be masked as satire or parody.

- *Interpretability:* Poor - concepts like *legitimate* or *artistic expression* are highly subjective. Needs clear, operational definitions and examples in the codebook.

- **Audience sufficiently clear to include people with ideological ties**
  - *Usability:* Medium - needs background research and may default to exclusion if ideology is unclear.
  - *Usefulness:* High when properly applied, as it helps target ideologically aligned incitement. However, overreliance on clarity of ideology can exclude meaningful cases.
  - *Interpretability:* Moderate - clear in extreme cases, ambiguous for users with mixed/fuzzy ideological markers. Needs clarification of what counts as *ideological*.

- **Deliberately creating an atmosphere of fear**
  - *Usability:* Moderate - can be applied, but with varying thresholds for what "atmosphere of fear" means.
  - *Usefulness:* Potentially useful for identifying incitement, but overlaps with other indicators. Subjectivity is acceptable if balanced by requiring multiple indicators.
  - *Interpretability:* Subjective and bias-prone - depends on the coder's perception of tone, threat, and emotional impact. Context-dependent.

- **Insults/incites hatred or violence**
  - *Usability:* Moderate to low - easy for explicit cases, difficult when disguised (irony, dog whistles, layered language).
  - *Usefulness:* High - it's a key indicator of problematic content, but overuse or misapplication can dilute its meaning.
  - *Interpretability:* Highly subjective. The scale from clearly offensive to subtly hateful is hard to standardise. Needs a spectrum approach or sub-categories.

- **Knowledge of or reference to discriminatory acts or speech**
  - *Usability:* Low - "knowledge of" is nearly impossible to prove unless the user admits it or has a clear history.
  - *Usefulness:* Limited - "reference to" is clearer and actionable; "knowledge of" is mostly speculative.
  - *Interpretability:* Poor - assessing internal knowledge is not feasible; focus should shift toward observable references only.

- **Deliberately creates the likelihood that a serious violent offence will be committed regardless of whether it is executed or not**
  - *Usability:* Low - deliberateness is extremely difficult to prove, especially online.
  - *Usefulness:* Theoretically crucial but practically risky - prone to overreach or underinclusion depending on coder bias.
  - *Interpretability:* Very subjective - requires assumptions about user intent, awareness, and downstream effects of speech. Highly influenced by the coder's worldview.

7.4.3 Characteristics of Implicit Extremist Content

To analyse which characteristics or combination of characteristics are needed to determine whether content qualifies as implicit extremist content, we build upon preliminary observations made during the data collection phase, existing research consulted for the literature review, and the team's own expertise in the field – a limited number of indicators to test the feasibility of a framework. These indicators have been selected based on their susceptibility to reveal (evasive techniques used to hide) the implicit harmfulness of the content, as follows:

- The **intent to conceal meaning (CM)** indicator captures elements that suggest an effort to obscure the 'true' meaning of the narrative put forward. This indicator and its scores/

variations have been identified based on the various textual and visual manipulation techniques observed during the data collection phase.

- The **harmful alliances/affiliations (HA)** indicator focuses on all signs denoting a connection to hateful or violent extremist groups or ideologies. This indicator and its scores/variations have been identified based on both preliminary observations made during the data collection phase and existing open-source databases documenting hateful and extremist symbols, emojis, and other coded language.

- **The problematic reference(s) to the historical/current context (PR)** indicator aim to situate the content within a broader context and historical narratives. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).
- The **implicit action trigger(s) (AT)** indicator identifies elements inserted in the narratives that might subtly encourage the audience to take actions. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

- The **presumed intent to cause harm (IH)** indicator considers potential harmful intentions behind content, focusing on *presumed* rather than definitive intent due to the inherent subjectivity in such assessments. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

Since a legal framework defining implicit extremist content is missing, there is no preset rule on how many indicators need to be present before such a qualification can be made. Whether the presence of one indicator in the post is enough to qualify the content as implicit extremist content, is related to the question of the threshold that needs to be met (see below). There are different options that can be considered. Given the loop function built into the pilot codebook, meaning that once implicit extremist content is detected, the post would need to be reassessed regarding presence of terrorist or illegal content, before a moderation decision can be taken in line with the principle respecting the freedom of expression, one can decide on the qualification if only one indicator is present, or rather opt for multiple indicators or a specific combination of indicators. However, the research team would strongly oppose only using one indicator, since it would most likely include too many posts that fall under the freedom of expression.

7.4.4 Detectability and Identifiability of Implicit Extremist Content

The indicators used were tested regarding their usability, usefulness, and interpretability. Of the 29 indicators used, six are considered to be usable, useful, and easy to interpret. There are some issues that need to be addressed with the 21 indicators.

More serious concerns arise regarding the following two indicators:

- **Instilling a Perceived Need for Self-Defence**
  - *Usability* Lack of clear operational criteria: The absence of concrete guidelines or illustrative examples makes it difficult to distinguish between legitimate expressions of concern (e.g., about safety or crime) and problematic narratives that incite fear for defensive action. Difficult to apply consistently: Without standardised thresholds, different coders may interpret similar content in diverging ways, reducing reliability.
  - *Usefulness:* The indicator captures a relevant rhetorical strategy often used to justify hostile or pre-emptive action. However, without clearer parameters, its practical value is limited and potentially counterproductive.

- *Interpretability:* Highly interpretative and prone to subjective judgment. This indicator relies heavily on individual perception, making it vulnerable to personal *or* ideological bias. What constitutes *self-defence* may vary significantly between coders, especially in politically charged contexts. Open to misuse or abuse. Coders or flaggers with specific agendas may apply this indicator selectively to content they disagree with, undermining the objectivity of the coding process.

- **Fostering Hate/Hostility Towards an Out-Group**
  - *Usability:* Requires clarification and concrete examples: Including a typology of in- and out-groups (e.g., religious, ethnic, political) and sample content would improve consistency and confidence in coding decisions.
  - *Usefulness:* Conceptually strong and highly relevant to identifying extremist content. However, its full potential depends on clearer guidance and calibration to avoid both under- and over-classification.
  - *Interpretability:* Ambiguity around the definition of *out-group*: It is unclear whether entities such as the government, elites, or institutions qualify as out-groups, especially when content criticises them from a populist or conspiratorial standpoint. Clearer definitional boundaries are needed. Blurring lines between political critique and harmful hostility: In-group/out-group dynamics are common in mainstream political discourse. Coders must assess not only who is being criticised but how - distinguishing between legitimate opposition and dehumanising rhetoric. Debatable thresholds for what qualifies as "fostering hate": Some content may express hostility or resentment without explicitly inciting hate or violence. Determining intent and impact in these cases is highly interpretive.

7.4.5 Key Challenges of Detection of Implicit Extremist Content and Recommendations for Use

There are five key overall concerns with the interpretability, usability, and potential of these indicators for inconsistent application in an assessment framework:

1. High subjectivity and risk of Bias
- This opens the door to subjective judgment and potential misuse by biased coders or flaggers, particularly when ideological agendas are involved.

2. Lack of Clear Definitions and Guidance
- There is a need for precise definitions (e.g., what constitutes an *out-group*) and clear thresholds for what type of content qualifies.
- The absence of concrete examples or typologies makes it hard for coders to apply the indicators consistently.

3. Blurry Boundaries Between Harmful and Legitimate Discourse
- Many expressions - such as political critique or discussion of societal threats—may fall into a grey area, making it difficult to determine if they meet the criteria for these indicators.
- There's a concern that ordinary or constructive discourse could be misclassified as extremist if not carefully distinguished.

4. Usability Undermined by Ambiguity
- Both indicators lack operational clarity, which challenges their practical use in a monitoring or moderation context.
- Coders may struggle to apply them in a reliable, consistent, and objective manner without additional training or resources.

5. Interpretability Issues Threaten Reliability
- The vagueness and reliance on contextual understanding undermine the reliability and inter-coder agreement, which are crucial for systematic content analysis.

For an accurate assessment framework, the indicators would thus benefit from improvements by clarifying:

- definitions and threshold,
- providing detailed guidance and examples across different ideological spectrums,
- addressing subjectivity and bias (for instance, by offering guidance on how to minimise personal bias, and by providing guidance and coder prompts),
- recognising that some ambiguity cannot be addressed yet offering recommendations on how to deal with that,
- operationalising complex indicators by providing decision-making frameworks or coding flowcharts,
- including triangulation with contextual clues to reduce over-reliance on the interpretation of one indicator, and
- by increasing consistency in application by providing example-based rules and exceptions, encouraging consensus-building practices or review protocols for borderline cases, and clearly stating what kinds of content should default to be excluded.

There are five key challenges identified regarding the usability, usefulness and interpretability of the indicators used to detect implicit extremist content:

1. High Subjectivity and Interpretative Variability
- Many indicators rely heavily on personal judgment and contextual interpretation, which introduces inconsistency.
- Examples include:
  - Whether something is satire, parody, or legitimate discourse.
  - Determining deliberate creation of fear or likelihood of violence.
  - Assessing whether content is insulting or incites hatred/violence, especially when concealed or coded.

2. Ambiguity and Lack of Clear Definitions
- Indicators like *group*, *legitimate*, *ideology*, or *capacity to commit violence* are vague or inconsistently understood, leading to coder disagreement.
- There is a need for clearer guidance or a refined codebook (e.g., what counts as a *legitimate* commemoration or contribution to public discourse).

3. Dependence on Contextual and Background Knowledge
- Accurate coding often requires in-depth knowledge of the user, audience, ideology, or platform context (e.g., subreddits, hashtags, user history).
- This reduces usability, as it makes the process time-consuming and difficult to scale or standardise.

4. Inconsistent or Low Thresholds for Validation
- Some indicators are validated too broadly or narrowly, depending on coder discretion (e.g., assuming all public posts target a group vs. narrowing it to clear group-directed posts).
- There is tension between erring on the side of caution and maintaining precision and reliability.

5. Limited Usefulness of Certain Indicators
- Some indicators (e.g., knowledge of discriminatory acts) are unclear in practical terms—hard to detect and difficult to prove intent or awareness.

While concerns or serious problems regarding usability could be addressed using regularly updated LLMs and databanks, in order to address the labour intensity of the assessment of the indicator, the serious problems with interpretability are not easily fixed. The interpretation of some of these indicators is highly subjective to personal bias, context, social, cultural or religious

acceptance, and furthermore might evolve over time. With the current ambiguity in terminology used in policy and legal frameworks, HSPs are offered insufficient guidance to be able to develop an accurate assessment framework. For an accurate assessment framework, the indicators would thus benefit from improvements by offering examples for the assessments of indicators, including thresholds for the scoring of indicators, further clarification of key terms, and more consideration of edge cases and coded expressions. It would, furthermore, be recommended that this is done on the basis of the four eyes principle, and by well versed professionals.

7.4.6 Cross-Platform and Cross-Ideology Comparison and Usability of a Codebook

When comparing the OSINT scores across platforms, TikTok stands out: nearly half of the content collected from TikTok was categorised as *implicit extremist content*. Notably, 97 percent of this content was associated with right-wing extremism.

In contrast, the presence of implicit extremist content was significantly lower on Instagram (14 percent) **and** Reddit (19.8 percent)**. The ideological distribution of content on these two platforms was also more diverse. On Instagram,** 48 percent of the content referenced unclear ideologies**,** 40 percent was linked to right-wing extremism, and 12 percent to Islamist ideologies. For Reddit, the majority of content (80 percent) referenced right-wing extremist ideologies, while the remaining 20 percent related to Islamist ideologies.

Looking at the classification of implicit extremist content by ideology, the differences are striking: 33.6 percent of *right-wing extremist* content was labelled as implicit extremist, compared to 14.6 percent of *unclear ideological content*, and only 11.8 percent of *Islamist* content.

This discrepancy suggests a few possibilities. Notwithstanding the possibility that in absolute numbers the supportive base for right-wing extremism is broader than for Islamist extremist ideology, thereby resulting in more right-wing extremist related posts in absolute numbers, it may indicate that current detection and moderation systems—whether automated or manual—are more effectively trained to identify implicit extremist content linked to Islamist or ambiguously ideological sources. Alternatively, it could reflect greater awareness or caution among the users of the platforms: Islamist extremists might be more conscious about what they place on these three platforms, or extreme right-wing users refer more to attempted concealing methods.

8. Findings, Challenges and Recommendations

Developing an accurate assessment framework for detecting terrorist, illegal, and implicit extremist content faces a series of entrenched challenges that limit both feasibility and reliability. Notably, these relate to definitional ambiguities, the lack of clear thresholds, and the heavy reliance on contextual interpretation, which makes classification, detection and moderation difficult. Distinguishing extremist rhetoric from satire, political critique, or ordinary expressions of grievance poses one of the biggest challenges. Their application often depends on extensive contextual knowledge and understanding the intent, which is difficult to grasp without knowing other kinds of content the user has posted on the same or other platforms, while coded or concealed expressions remain especially resistant to consistent classification. Further obstacles include the evolving tactics of extremist actors, the lack of indicators that are applicable across different ideologies, the limited practical usefulness of certain indicators, and the overwhelming scale of online content, which exceeds the capacity of both human coders and current automated systems. Together, these weaknesses undermine reliability, increase the risk of misclassification, threaten the credibility of any assessment framework built on such indicators, and ultimately unduly infringe on fundamental rights of users, rendering the feasibility of a reliable assessment framework at the current state of the policies highly unlikely.

Despite these obstacles and given the fact that our open-source intelligence (OSINT) research still yielded plenty of content that would merit moderation, improvement of existing detection mechanisms used by hosting service providers (HSPs) to capture implicit extremist content is highly recommended and should be possible through more targeted and adaptive strategies. Indicators can be strengthened by refining definitions and by offering benchmarks catering to different ideologies and guiding examples, developing typologies of in- and out-groups, and providing detailed examples that help distinguish legitimate critique or humour from harmful rhetoric. And although not much is known about the detection mechanisms of HSPs due to the lack of transparency, news reports suggest that there is a trend of limiting the staff involved in human assessment, leaving it now mostly to automated detection systems. This trend contrasts with one of the key findings of this study, which suggests that assessment of various indicators is highly context-specific, needing a human as opposed to automated assessment. This is particularly the case with implicit extremist content, where the interpretation of concealed meanings and context plays a significant role in the classification of the content. Furthermore, the usability of existing detection mechanisms would benefit from operationalising complex indicators with coding flowcharts, decision-making frameworks, and triangulation with contextual clues, while subjective biases can be mitigated through coder training, review protocols, and consensus-building practices. Assessment frameworks used for the detection of certain online content should be treated as “living documents” that evolve with emerging tactics, supported by domain-specific and culturally sensitive approaches rather than one-size-fits-all frameworks. To maintain the “living document”, multi-stakeholder collaboration - bringing together researchers, practitioners, platforms, civil society, and affected communities - would be essential to balance accuracy with safeguards against misuse. Coordination and information exchange between HSPs on how they deal with these issues would also contribute to keeping assessment frameworks up to date. A hybrid model that combines AI-assisted pre-screening with human expertise can help manage the sheer amount of content while ensuring contextual sensitivity. It would, however, be vital that the role of the human assessment is maintained in a meaningful manner, and sufficient human moderators are kept on the payroll to assess the category of content where contextual interpretation is key. In this way, while a universal solution is not feasible, taking into account a clear set of conditions of how to work with indicators could at least strengthen detection while protecting legitimate expression.

The question of whether it is feasible to develop a codebook for detecting implicit extremist content online thus presents significant legal, ethical, and practical challenges. This chapter synthesises key observations and findings based on the research conducted for this study to assess the feasibility of such an undertaking. It answers the key research questions as laid out in chapter 2, and offers several ethical considerations related to the desirability of an assessment framework.

Research Questions

- **Research question 1: What features or combination of features in online content determine whether it constitutes extremist or terrorist content on online platforms?**

- **Research question 2: In what ways and to what extent can these features—either individually or in combination—be detected and identified in online platform content?**

Analysis of the legal frameworks yielded relevant elements for the detection of terrorist and illegal content to be used in a potential interpretative framework. In addition, existing research combined with observations from the OSINT research provided a set of indicators that can be used for the detection of implicit extremist content:

- The **intent to conceal meaning (CM)** indicator captures elements that suggest an effort to obscure the 'true' meaning of the narrative put forward. This indicator and its scores/variations have been identified based on the various textual and visual manipulation techniques observed during the data collection phase.

- The **harmful alliances/affiliations (HA)** indicator focuses on all signs denoting a connection to hateful or violent extremist groups or ideologies. This indicator and its scores/variations have been identified based on both preliminary observations made during the data collection phase and existing open-source databases documenting hateful and extremist symbols, emojis, and other coded language.[433]

- The **problematic reference(s) to the historical/current context (PR)** indicator aim to situate the content within a broader context and historical narratives. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

- The **implicit action trigger(s) (AT)** indicator identifies elements inserted in the narratives that might subtly encourage the audience to take actions. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

- The **presumed intent to cause harm (IH)** indicator considers potential harmful intentions behind content, focusing on *presumed* rather than definitive intent due to the inherent subjectivity in such assessments. This indicator and its scores/variations have also been informed by prior ICCT research examining violent extremist disinformation (not publicly available).

The elements of terrorist, illegal, and implicit extremist content were included in the pilot codebook presented in chapter 7. Per category of content (terrorist, illegal, or implicit extremist content), it was made clear whether all the elements needed to be met, or whether a single, or combination, of elements needed to be met for the content to qualify as either one of the three

433 See e.g. ADL, *Hate On Display™*; "Global Extremist Symbols Database," Global Project Against Hate and Extremism; International Centre for Digital Threat Assessment, *Interpreting and Translating Emojis*.

categories. Yet, while determination of the key indicators is possible, the use of the indicators themselves is not without challenges.

Definitional Challenges and Conceptual Ambiguity

The Problem of Undefined Terminology

A fundamental obstacle lies in the lack of clear, internationally accepted definitions of core concepts. The absence of universally agreed-upon definitions for terrorism, violent extremism, hate speech, and borderline content creates significant challenges for platforms, researchers, and practitioners attempting to address the scale of harmful content. Research has identified this definitional vacuum as a crucial hurdle in allowing for a human rights compliant content moderation across different platforms, particularly given the lack of consensus on what constitutes *hate speech* or *harmful content*.[434]

While the DSA references "illegal hate speech" as content requiring moderation with "consideration to the impact of such measures on fundamental rights", no universally accepted definition exists to operationalise this requirement.

The Complexity of Indirect Incitement and Glorification

Even where terrorist offences and illegal content are more clearly defined, the concepts of direct and indirect incitement and glorification remain poorly delineated. The UN Human Rights Committee has emphasised that offences such as "encouragement of terrorism" and "extremist activity," as well as those involving "praising," "glorifying," or "justifying terrorism," should be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression."[435] Similarly, to the vague and broad definitions of hate speech, illegal hate speech, and harmful content, this definitional challenge also impacts the feasibility of creating automated detection criteria, as broad interpretations risk unlawful interference with fundamental rights. Currently, a Bill is pending in the Netherlands related to the criminalisation of glorifying terrorism, including when this is done online, which is being criticised for potentially restricting lawful political speech and merits careful revision.

The Borderline Content Conundrum

So-called borderline content presents particular challenges for the development of an assessment framework due to its nature as an umbrella term encompassing different categories of content, ranging from self-harm to disinformation to gore content. The focus of this study is to look at borderline content in the context of terrorism and violent extremism, therefore using the term 'implicit extremist content'. Platforms, however, often use the term 'borderline', and characterise borderline content as material that may not yet violate their policies but is concerning. At the same time, regulatory bodies and law enforcement have a different understanding of what constitutes borderline content. Since borderline content can relate to multiple policy areas beyond extremism and terrorism, a one-size-fits-all approach seems difficult to implement by all relevant stakeholders.

HSPs encounter significant difficulties in implementing uniform approaches to detect implicit extremist content due to divergent norms and values across different communities and cultures. Implicit extremist content that may be deemed harmless in one context can be perceived as harmful in another, with these contextual interpretations evolving over time. This cultural relativity poses a substantial challenge to developing universal detection criteria, as any assessment framework

434 ARTICLE 19, *Content moderation and freedom of expression handbook*, 8.
435 General comment No. 34, 102 UN Human Rights Committee.

would need to account for contextual nuances that vary significantly across geographical, cultural, and temporal dimensions.

Diversity in HSP's policies

Related to the previous obstacle is the fact that HSPs are first and foremost private actors, and their primary objective is generating revenues by maintaining a distinctive profile, which plays a role in targeting a preferred audience of users. This objective has implications for the way they perceive and qualify certain content to be in line with their Terms of Use (ToS) or not. Their interest in developing a universal assessment framework might therefore be minimal.

- **Research question 3: What can be said about the reliability of detection methods for extremist and terrorist content on online platforms?**

- **Research question 4: How does the reliability of detection methods for various types of harmful content on online platforms (terrorist, clearly extremist, and borderline) relate to the risk of incorrect moderation decisions by online platforms, which may result in violations of the fundamental right to freedom of expression?**

From the coding of content done for this study, it can be concluded that of the 26 indicators assessed (some are used for different acts) to identify terrorist or illegal content, ten were found to be usable, useful, and easy to interpret. Eight required adjustments, and another eight raised more serious concerns.[436] The problematic indicators share issues of ambiguity, subjectivity, and heavy dependence on context, making them difficult to apply consistently. Terms such as "group," "legitimate," or "deliberately" are often vague and open to interpretation, leading to inter-coder disagreement and inconsistent application. Several indicators (e.g., audience clarity, atmosphere of fear, insults/incitement) are conceptually important but prone to bias, overlap, or over-inclusion without stricter definitions. Others (e.g., knowledge of discriminatory acts, awareness of audience capability, deliberate creation of violent likelihood) are practically unworkable, as they rely on assumptions about user intent, awareness, or internal knowledge. Overall, while some indicators remain central to identifying harmful content, their current formulations require clearer operationalisation, definitions, and examples to ensure reliability and avoid misapplication.

In our assessment, testing the 29 potential indicators to identify implicit extremist content, only six were deemed usable, useful, and easy to interpret, while 21 showed issues requiring attention. Two indicators raised particularly serious concerns. *Instilling a perceived need for self-defence* lacks clear operational criteria and thresholds, making it difficult to distinguish legitimate safety concerns from problematic incitement, and leaving it highly interpretive, subjective, and open to bias or misuse. *Fostering hate/hostility towards an out-group* is conceptually important but undermined by definitional ambiguities and unclear thresholds—particularly in distinguishing harmful hostility from legitimate political critique. Both indicators capture key dynamics of extremist discourse but require clearer definitions, examples, and boundaries to ensure consistent, objective, and reliable application.

436 The eight indicators requiring adjustment were:

Table 8: Combined usability, usefulness, and interpretability scores for indicators

INDICATORS	Usability / usefulness / interpretability
Crosscutting and similar indicators	
Intended audience is a **group** (IT/IV)	
Content is spread in **public** (IT/HS/IV/DIC)	
Awareness of the likelihood that the audience could commit a terrorist act / serious violent offence - **capacity** (IT/IV)	
Knowledge of or reference to terrorist activities (IT/GT)	
Not a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (IT/GT/RT/HS/IC/DIC)	
Audience sufficiently clear to include people with **ideological ties** (IT)	
Directly or indirectly addressing an audience - individual or group (RT)	
Incitement to commit/participate in a terrorist offence (IT)	
Advocate the commission of a terrorist offence (IT)	
Deliberately creates the likelihood of a **serious** terrorist offence regardless of whether executed or not (IT)	
Direct or indirect **causal link** between incitement and terrorist act (IT)	
Imminent risk that a terrorist act will be committed (IT)	
Glorification of a terrorist act (GT)	
Justification of past/future serious terrorist offence (GT)	
Deliberately creating an **atmosphere of fear** (GT)	
Recruitment for a terrorist organisation (RT)	
Call to supply information/material resources/funding/human resources to a terrorist group (RT)	
Knowledge of or reference to a specific terrorist group, one or more serious criminal acts committed by the referenced terrorist group, one or more concrete aims of the terrorist group (RT)	
Hate speech (HS)	
Discriminates against a group or individual (HS)	
Discriminates **on the basis of** race, gender, religion, belief, sexual orientation, physical or mental disability (HS)	
Insults/incites hatred or violence (HS)	
Knowledge of or reference to discriminatory acts or speech (HS)	
Incitement to violence (IV)	
Direct and imminent causal link between incitement and serious violent offence (IV)	
Advocates commission of a serious violent offence (IV)	
Deliberately creates the likelihood that a **serious** violent offence will be committed regardless of execution or not (IV)	
Knowledge or reference to serious acts of violence (IV)	
Denial, downplaying or justification of international crimes (DIC)	
Against a **group or individual** (DIC)	
Denies, justifies, or downplays core international crimes established irrevocably by a court (DIC)	
Knowledge or reference to the fact that crimes constitute core international crimes (DIC)	
Intent to conceal meaning (CM)	
Ambiguous or coded language (CM)	
Altered text or misspelling (CM)	
Misleading cover image for video (CM)	

Blurred or altered image - e.g. marker function to hide certain words (CM)	
Use of humour or irony (CM)	
Harmful alliances/affiliations (HA)	
Hateful / extremist symbols (RWX)	
Hateful / extremist symbols (Islamist)	
Hateful / extremist emojis (RWX)	
Hateful / extremist emojis (Islamist)	
Hateful / extremist coded slogans, slang, or acronyms (RWX)	
Hateful / extremist coded slogans, slang, or acronyms (Islamist)	
Problematic references to historical/current context (PR)	
Denial or questioning of proven past or present crimes (PR)	
False or misleading claims about past or present crimes (PR)	
Justification of current or potential future crimes through references to past crimes (PR)	
Falsified historical claims aimed at denying the existence or territorial legitimacy of a state (PR)	
Glorification or positive portrayal of individuals or groups involved in crimes (PR)	
Glorification or positive portrayal of public figures known for spreading hateful or extremist narratives (PR)	
Promotion or endorsement of books, films, essays, or any other products known for spreading or supporting any of the above narratives (PR)	
Implicit action triggers (AT)	
Instilling a perceived need for retaliation (AT)	
Instilling a perceived need for self-defence (AT)	
Instilling a perceived need to protect a group from reputational damage (AT)	
Instilling the idea of compensation or reward for undertaking action (AT)	
Exerting peer pressure/respect for a code of honour (AT)	
Providing sacred justification for action (AT)	
Presumed intent to cause harm (IH)	
Normalise hateful/violent narratives (IH)	
Propagate hateful/violent conspiracies (IH)	
Foster hate/hostility towards an out-group (IH)	
Trigger to take online harmful actions - e.g. harassment/doxxing (IH)	
Trigger to take offline harmful actions (IH)	

Detection Challenges and Technological Limitations

The literature review, the outcomes of the expert meeting, and the interviews conducted further indicated that detection and content moderation systems face increasing sophistication in concealment tactics employed by users seeking to evade detection. A particular concern arises when users employ humour, coded language or memes understood only by specific groups or communities, thereby evading detection by content moderation systems despite the content being harmful or unlawful. As this content remains visible online, it can reach broader audiences and support the growth of extremist communities.

The sheer volume of online content, combined with the increasing use of emerging technologies and evolving concealment tactics, creates practical challenges for HSPs in effectively detecting and moderating content. Determining the nature of content is becoming increasingly complex, and as a result, HSPs often fail to identify certain content as hate speech or other forms of illegal content. These practical limitations suggest that any assessment framework would need to account for scalability challenges and the limitations of current detection technologies.

The Role of Private Actors in Upholding Fundamental Rights Online

Scholars acknowledge that HSPs play a crucial role in governing public discourse online, holding a trove of users' data required for content moderation and being in charge of the spaces in which public discourse takes place online.[437] Nevertheless, considerable disagreement arises about how much leeway these private, profit-oriented actors should be given in deciding on normative frameworks that ultimately shape public discourse and on which infringements on the right to freedom of speech are being made.

Notably, these private actors have no general obligation to ensure individuals' enjoyment of human rights, as this obligation primarily lies with the States.[438] Yet there is growing recognition that HSPs should abide by the UN Guiding Principles on Business and Human Rights.[439] Acknowledging the potential of human rights violations by HSPs in content moderation, the EU took a regulatory approach that aims to provide common definitions and uniform approaches at least on a regional level and seeks to increase transparency over the actions taken by HSPs in content moderation and the assurance of minimum human rights safeguards, such as effective remedies.[440]

However, reports indicate that many platforms have failed to achieve an appropriate balance between safeguarding freedom of expression and enforcing content moderation policies.[441] The right to freedom of expression appears critically threatened on the internet as outlined above. Additionally, content moderation, if not carefully managed, can disproportionately affect the right to freedom of expression of minority, marginalised, or activist groups when content targeting these groups is removed, as content moderation mechanisms may include biases related to sexual orientation, race, gender, or religion, which could lead to discrimination against specific groups.[442] Ultimately, courts of law play a pivotal role in adjudicating cases concerning content moderation and the right to freedom of expression, thereby providing guidance for HSPs on what measures to take in order to protect the freedom of expression and other fundamental rights in the context of content moderation.[443] However, such proceedings tend to be protracted and only pursued in a very limited number of cases, which delays the effective protection of the right to freedom of expression by judicial bodies.

The question is whether the implementation of a universal assessment framework will contribute to a more responsible guardianship by HSPs of the freedom of expression, or whether they will perceive it as a framework that increases the risk of not complying with high fines as a potential consequence remains difficult to assess, as HSPs primary objective is running a business and are guided by other principles.

- **Research question 5: With current knowledge and technological capabilities, is it possible to develop a valid and reliable interpretive framework for detecting and identifying both explicit extremist and terrorist content, as well as more implicit borderline content on online platforms, without unjustly infringing upon the fundamental right to freedom of expression?**

- **Research question 6: Under what conditions could the implementation of an interpretive framework for extremist and terrorist online content contribute to reducing both the dissemination of such harmful content and the potential radicalisation of internet users?**

437 Dvoskin, "The Illusion of Inclusion," 1317 & 1332-1333.
438 Jørgensen, "When private actors govern human rights," 346-363.
439UN Human Rights Council. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. New York & Geneva: UN, 2011. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf
440 European External Action Service, *EU Guidelines on Freedom of Expression Online and Offline*, 1; §§ 6 & 40 TCO.
441 David Kaye, *Promotion and protection of the right to freedom of opinion and expression*, 16.
442 Wolfgang Schulz et al., *Algorithms and human rights*, 27-28.
443 Tuchtfeld, "Case law on content moderation and freedom of expression," 8.

If a valid and reliable interpretive framework is deemed unfeasible:

- **Research question 7: What are the legal, technical, and possibly other barriers that hinder the development of an interpretive framework for extremist online content? What conditions must be met to remove these barriers, and how feasible is this?**

The analysis highlights that while indicators are central to identifying implicit extremist content, their current formulation faces deep limitations that undermine reliability and consistent application. These challenges arise not only from the inherent complexity and subjectivity of implicit communication but also from structural issues such as vague definitions, cultural relativity, and the evolving strategies of extremist actors. These speedy developments would therefore require constant updates of any framework. Contextual elements – according to all interviewees – are furthermore overall considered to be hard to operationalise. Unless addressed, these weaknesses risk misclassification, bias, reduced credibility of any assessment framework, and bear the potential to severely infringe on the fundamental rights of users, while effective remedies are not always available to them.

Regulatory and Implementation Challenges

Both the TCO Regulation and the DSA have only recently entered into force, making it difficult to assess their main challenges and benefits and whether their interpretation by HSPs may lead to false positives or false negatives in detecting and moderating relevant content. As the practical implications of current legal frameworks and practical implementation remain unclear without meaningful observations that could inform, this also complicates the development of automated detection approaches.

The terms of use, community guidelines, and policies of platforms are not easily accessible, making it difficult for users to understand what is permissible and how to challenge moderation decisions. This lack of transparency regarding detection and moderation has resulted in limited research on users' perceptions and experiences with these actions, notably regarding their fairness. Such opacity would complicate the development and validation of any automated assessment framework.

Technological Opportunities and Limitations

Larger platforms, as far as they also deploy human content moderators, already make use of AI tools to support moderators in making complex contextual judgments, especially when classification-based tools in content moderation are primarily used to flag content for further human review. Regular evaluation of this approach and transparency on its effectiveness, however, lags behind. Although limited research has focused on the effectiveness of the application of machine learning in detecting radical behaviour and terrorist activities online, its successful use in other domains suggests it could yield similarly effective results in identifying terrorist, illegal and implicit extremist content online, provided these evaluations are made public.
Evidence indicates that individuals who were notified of their violation of platform community standards, accompanied by links to relevant rules, were more likely to comply with those standards.[444] This suggests that transparent, well-designed detection systems could enhance compliance while maintaining user rights, potentially supporting the implementation of codebook-based approaches.

444 Katsaros et al., "Online Content Moderation,» 70.

Challenges, Ethical Considerations and Recommendations

Challenges

The main challenges in developing an accurate and reliable assessment framework can thus be summarised as follows:

- **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.
  - Distinguishing extremist rhetoric from legitimate freedom of expression, such as satire, political critique, or legitimate debate, is highly challenging.
  - 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.

Ethical Considerations

In addition to the legal, practical and technical considerations addressed above, there are also several ethical considerations that should not be overlooked in the discussion. Efforts to develop and apply indicators for detecting terrorist, illegal and implicit extremist content raise significant ethical challenges that, after all, cannot be addressed through technical refinements alone. This was also extensively discussed during the expert meeting. The risks of misclassification, bias, overreach, and misuse call for a framework that is anchored in ethical theory as well as human rights standards. Since we already addressed the human rights considerations, we will only briefly address how human rights interact with a duty-based ethics lens. Below, we thus offer

several ethical lenses and considerations based on duty-based, consequentialist, virtue, and justice-oriented approaches, to assess the indicators and practical use.

Duty and Rights-Based Ethics

A deontological and human rights perspective underscores the duty to protect fundamental freedoms, particularly freedom of expression (ICCPR Article 19), privacy, and due process.[445] In accordance with this approach, indicators must therefore be clearly defined, transparent, and applied in ways that avoid arbitrariness or hidden criteria. Ambiguity around terms undermines this duty, as does the absence of safeguards against state misuse for silencing dissent. A rights-based approach requires strict adherence to principles of legality, necessity, and proportionality: indicators should only restrict speech when clearly justified to prevent tangible harm.

Consequentialist Ethics

From a utilitarian perspective, the most influential form of consequentialism, indicators are ethically justifiable only insofar as they minimise harm (e.g., the spread of extremist rhetoric inciting violence) while avoiding disproportionate costs (e.g., chilling effects on public debate).[446] This requires careful calibration of thresholds and a willingness to reassess tools based on their demonstrated impact. For example, over-inclusive detection may protect vulnerable groups but risks stifling political discourse; under-inclusive approaches may safeguard expression but allow harmful content to circulate. Ethical use thus requires systematic impact assessments that weigh these trade-offs transparently.

Virtue Ethics

ndicator use is not only a technical process but also a matter of cultivating professional virtues among those who design and apply them.[447] Coders, moderators, and analysts must be trained to exercise prudence, humility, and fairness, particularly when working in highly interpretative contexts. Overconfidence or ideological bias in coding risks undermining trust and legitimacy. Multi-stakeholder involvement also reflects the virtue of *practical wisdom (phronesis)*, ensuring that diverse perspectives inform judgments and reduce the likelihood of systematic blind spots.

Ethics of Care

Beyond abstract principles, an ethics of care, by some considered a sub-category of virtue ethics, highlights the relational and contextual nature of extremist discourse. What may appear harmful in one cultural or political context may not be in another. Frameworks should therefore prioritise empathy and attentiveness to the lived experiences of both those harmed by extremist rhetoric and those at risk of unjust restriction.[448] This includes designing culturally sensitive indicators and embedding safeguards against erasing legitimate political critique under the guise of security.

Justice and Fairness

A Rawlsian perspective emphasises procedural fairness and protection for the least advantaged. Indicators should be developed and applied in ways that do not disproportionately burden minorities, dissidents, or vulnerable communities.[449] This implies inclusive participation in assessment framework development, transparent review mechanisms, and recourse for

445 Patrick Overeem, "Public service ethics and integrity: Some general lessons," in *Ethics of counterterrorism* (Meppel: Boom, 2017), 22.
446 Overeem, "Public service ethics and integrity," 22.
447 Overeem, "Public service ethics and integrity," 23.
448 Virginia Held, "The Ethics of Care," in *The Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press, 2009), 537-566.
449 Salla Westerstrand, "Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence," *Science and Engineering Ethics* 30, no. 46 (2024), https://link.springer.com/article/10.1007/s11948-024-00507-y?trk=public_post_comment-text.

individuals or groups harmed by misclassification. Justice also demands consistency: similar cases must be treated alike, requiring clearer operational guidance, typologies, and coder protocols to reduce interpretive variability.

Recommendations

Conditions for a reliable and accurate assessment framework

While the challenges identified above are significant, the findings also point toward strategies that can strengthen both the reliability and usefulness of indicators and that can assist in supporting their usability through overall improvement of clarity, adaptability, and collaboration. This study and the pilot codebook could furthermore be used to improve current detection and moderation approaches of HSPs, and guide the ongoing public-private cooperation and dialogue in an effective and rule of law compliant manner to manage terrorist, illegal, and implicit extremist content online. The recommendations below 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**
   - Abide by the obligation to implement precise definitions for vague terms;
   - Establish thresholds for incitement, hate, or hostility that do not restrict the freedom of expression;
   - Be clear on the combination of indicators that need to be met to qualify content as implicit extremist content.

2. **Guidance and Examples**
   - Offer illustrative examples across ideological spectrums, including examples of when content is protected by the freedom of speech, such as satire, critique or harmful but lawful content.
   - Use decision-making frameworks or coding flowcharts to standardise the application.
   - Develop typologies of in-/out-groups to guide application.

3. **Subjectivity and Bias Reduction**
   - Provide coder training, prompts, and bias-mitigation strategies.
   - Adopt “four-eyes” review protocols and consensus-building practices for borderline cases.

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

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

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

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

8. Transparency and Appeal
- Improve transparency and reporting on moderation decisions.
- Need for proportionality of 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.
- Provide clear information on how and where a moderation decision can be appealed.

Specific Recommendations for policymakers:

This study, commissioned by WODC, 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, 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 NCTV and the ATKM.

Policymakers are recommended to:

1. **Refrain from using the term 'borderline' content**, as that will only further contribute 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 need to be improve in 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**
   e. On what is harmful and unlawful, and what is harmful but lawful content;
   f. 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 it 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:**
   g. To conduct an open discussion on the indicators they use in their assessment frameworks and whether they use different assessment frameworks for different ideologies;
   h. 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.

Concluding Remarks

A universal, valid and reliable assessment framework for detecting implicit extremist content does not appear feasible. However, more reliable and adaptable frameworks used by HSPs might be achievable. By refining definitions, operationalising complex indicators, and embedding iterative learning, hybrid systems, and collaborative oversight, used indicators can be transformed into more consistent and credible 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.

Bibliography

Adams, Steve. “A Guide to OSINT Investigation and Research on TikTok.” Published November 13, 2019. https://www.intelligencewithsteve.com/post/osint-tiktok.

ADL. Hate On Display™: *Hate Symbols Database*. New York: ADL, 2019. https://www.adl.org/sites/default/files/ADL%20Hate%20on%20Display%20Printable_0.pdf.

ADL Center on Extremism. “Steam-Powered Hate: Top Gaming Site Rife with Extremism & Antisemitism.” *ADL* (November 2024). https://www.adl.org/resources/report/steam-powered-hate-top-gaming-site-rife-extremism-antisemitism.

AIVD. *A web of hate: The online hold of extremism and terrorism on minors.* The Hague: Ministry of the Interior and Kingdom Relations, 2025. https://english.aivd.nl/publications/publications/2025/04/03/a-web-of-hate.

AIVD. *Anti-institutional extremism in the Netherlands*: *A serious threat to the democratic legal order?* The Hague: Ministry of the Interior and Kingdom Relations, 2023). https://english.aivd.nl/binaries/aivd-en/documenten/publications/2023/11/7/publicatie-anti-institutional-extremism-in-the-netherlands/Publicatie-Anti-institutional-extremism-in-the-Netherlands-ENG.pdf.

AIVD. “Wat is het verschil tussen activisme, extremisme en terrorisme?” Accessed September 10, 2025. https://www.aivd.nl/onderwerpen/extremisme/vraag-en-antwoord/wat-is-het-verschil-tussen-activisme-extremisme-en-terrorisme.

Albert, John. “DSA risk assessment reports: A guide to the first rollout and what’s next.” Published December 9, 2024. https://dsa-observatory.eu/2024/12/09/dsa-risk-assessment-reports-are-in-a-guide-to-the-first-rollout-and-whats-next/.

Albrecht, Stephen, and Merle Strunk. „Memes für die Massen: Rechtspopulistische Fake- Accounts und ihre visuellen Strategien.“ In *Bilder, soziale Medien und das Politische: Transdisziplinäre Perspektiven auf visuelle Diskursprozesse*. Bielefeld: transcript Verlag, 2021. https://doi.org/10.1515/9783839450406-007.

Alkiviadou, Natalie. “Platform liability, hate speech and the fundamental right to free speech.” *Information & Communications Technology Law* 34, no. 2 (2024). https://doi.org/10.1080/13600834.2024.2411799.

Allchorn, William. “Turning Back to Biologised Racism: A Content Analysis of Patriotic Alternative UK’s Online Discourse.” *GNET Insights* (2021). https://gnet-research.org/2021/02/22/turning-back-to-biologised-racism-a-content-analysis-of-patriotic-alternative-uks-online-discourse/.

Appeals Centre Europe. “Rules of Procedure.” Accessed March 11, 2025. https://www.appealscentre.eu/rules-of-procedure/.

Appeals Centre Europe. “Startpagina.” Accessed March 11, 2025. https://www.appealscentre.eu/.

Appelman, Naomi, João Pedro Quintais, and Ronan Fahy. “Article 12 DSA: Will platforms be Required to apply EU fundamental rights in content moderation decisions?” Published December 9, 2024. https://dsa-observatory.eu/2021/05/31/article-12-dsa-will-platforms-be-required-to-apply-eu-fundamental-rights-in-content-moderation-decisions/.

Are, Carolina. "'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation." *Information, Communication & Society* (2024). https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2396621.

Arora, Arnav, Preslav Nakov, Momchil Hardalov et al. "Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go." *ACM Computing Surveys* 56, no. 3 (2023). https://doi.org/10.1145/3603399.

ARTICLE 19. *Content moderation and freedom of expression handbook*. London: ARTICLE 19, 2023. https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf.

ARTICLE 19. *Contents Social Media Councils: One piece in the puzzle of content moderation*. London: ARTICLE 19, 2021. https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf.

Askanius, Tina. "On Frogs, Monkeys, and Execution Memes: Exploring the Humor-Hate Nexus at the Intersection of Neo-Nazi and Alt-Right Movements in Sweden." *Television & New Media* 22, no. 2 (February 2021). https://doi.org/10.1177/1527476420982234.

Askanius, Tina, Bàrbara Molas, and Amarnath Amarasingam, "Far-right extremist narratives in Canadian and Swedish COVID-19 protests: a comparative case study of the Freedom

Movement and Freedom Convoy." *Behavioral Sciences of Terrorism and Political Aggression* 17, no. 2 (2024). https://doi.org/10.1080/19434472.2024.2340492.

Baele, Stephane, Elahe Naserian, and Gabriel Katz. "Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey." *Terrorism and Political Violence* (2024). https://doi.org/10.1080/09546553.2024.2380089.

Broekaert, Clara, Colin Clarke, Michaela Millender, Annika Scharnagl, and Joseph Shelzi. "Accelerating Hate: The Impact of October 7 on Terrorism and Political Violence in the West." *The Soufan Center* (September 2024). https://thesoufancenter.org/wp-content/uploads/2024/10/TSC-Special-Report-Accelerating-Hate-The-Impact-of-October-7-on-Terrorism-and-Political-Violence-in-the-West.pdf.

Bronzwaer, Stijn. "Complete moderatieteam TikTok in Nederland ontslagen." *NRC*. Published October 14, 2024. https://www.nrc.nl/nieuws/2024/10/14/complete-moderatieteam-tiktok-in-nederland-ontslagen-a4869230.

Brown, Alexander. *Models of Governance of Online Hate Speech: On the emergence of Collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe*. Strasbourg: Council of Europe, 2020. https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d.

Busch, Ella, and Jacob Ware. "The Weaponisation of Deepfakes: Digital Deception by the Far Right." *ICCT* (December 2023). https://www.icct.nl/sites/default/files/2023-12/The%20Weaponisation%20of%20Deepfakes.pdf.

Buttarelli, Giovanni. *Formal comments of the EDPS on the Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online*. Brussels: EDPS, 2019. https://www.edps.europa.eu/sites/default/files/publication/2018-02-13_

edps_formal_comments_online_terrorism_regulation_en.pdf.

Carpani, Anna Maria, Tom Siegel, Ray Liu et al. *EU Internet Forum: Study on the Role and Effects of the Use of Algorithmic Amplification to Spread Terrorist, Violent Extremist and Borderline Content*. Luxembourg: Publications Office of the EU, 2023.

CDMSI. *CONTENT MODERATION: Best practices towards effective legal and procedural Frameworks for self-regulatory and co-regulatory mechanisms of content moderation*. Strasbourg: Council of Europe, 2021. https://edoc.coe.int/en/internet/10198-content-moderation-guidance-note.html.

CQCore, and Roshan Behera. "The-Osint-Toolbox / Social-Media-OSINT." GitHub. Accessed January 21, 2025. https://github.com/The-Osint-Toolbox/Social-Media-OSINT?tab=readme-ov-file#faceboo.

Doody, Sean, and Michael Jensen. "Hash-Sharing Database Review: Challenges and Opportunities." *GIFCT Year 4 Working Group* (December 2024). https://gifct.org/wp-content/uploads/2025/02/GIFCT-24WG-1224-HSDR-Challenges-1.1.pdf.

Douek, Evelyn. "Content moderation as systems thinking." *Harvard Law Review* 136, no. 2 (2022). https://harvardlawreview.org/wp-content/uploads/2022/11/136-Harv.-L.-Rev.-526.pdf.

Douek, Evelyn. "The Meta Oversight Board and the Empty Promise of Legitimacy." *Harvard Journal of Law & Technology* 37, no. 2 (2024). https://jolt.law.harvard.edu/assets/articlePDFs/v372/3-The-Meta-Oversight-Board-and-the-Empty-Promise-of-Legitimacy.pdf.

Drolsbach, Chiara, and Nicolas Pröllochs. "Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database." *WWW '24: Companion Proceedings of the ACM Web Conference* 2024 (2024). https://doi.org/10.1145/3589335.3651482.

Dvoskin, Brenda. "The Illusion of Inclusion: The False Promise of the New Governance Project For Content Moderation." *Fordham Law Review* 39, no. 4 (2025). https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=6141&context=flr.

Eder, Niklas. "Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop To Address the Societal Harms of Content Moderation." *German Law Journal* (2024). https://doi.org/10.1017/glj.2024.24.

Einwiller, Sabine, and Sora Kim. "How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation." *Policy & Internet* 12, no. 2 (2020). https://doi.org/10.1002/poi3.239.

Enarsson, Therese. "Navigating hate speech and content moderation under the DSA: insights from ECtHR case law." *Information & Communications Technology Law* 33, no. 3 (2024). https://doi.org/10.1080/13600834.2024.2395579.

eSafety Commissioner. "The Global Online Safety Regulators Network." Accessed September 10, 2025. https://www.esafety.gov.au/about-us/consultation-cooperation/international-engagement/the-global-online-safety-regulators-network.

EU Agency for Fundamental Rights. *Online content moderation - Current challenges in*

detecting hate speech. Luxembourg: Publications Office of the EU, 2023. https://fra.europa.eu/en/publication/2023/online-content-moderation.

EU Internet Forum. *EU Internet Forum at 10 YEARS: Celebrating the achievements of the first decade's cooperation to fight harmful and illegal content online*. Brussels: European Commission, 2024. https://home-affairs.ec.europa.eu/document/download/0fb0be8a-c145-4948-a260-0a6be11ffc53_en?filename=EU%20Internet%20Forum%20Brochure.pdf&prefLang=uk.

European Court of Human Rights. *Factsheet - Hate speech*. Strasbourg: Council of Europe, 2023. https://www.echr.coe.int/documents/d/echr/FS_Hate_speech_ENG.

European Parliament, and Ipsos European Public Affairs. "Youth Survey 2024." *FlashEurobarometer* February 2025). https://europa.eu/eurobarometer/surveys/detail/3392.

European Commission. "Call for participants: RAN C&N meeting on How to deal with Borderline content (related to hate speech, meme culture, humour etc.) from the perspective of public trust?" Published March 1, 2024. https://home-affairs.ec.europa.eu/news/call-participants-ran-cn-meeting-how-deal-borderline-content-related-hate-speech-meme-culture-humour-2024-03-01_en.

European Commission. *Code of Conduct on Countering Illegal Hate Speech Online*. Brussels: European Commission, 2016. https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_en.

European Commission. *Code of Conduct on Countering Illegal Hate Speech Online +*. Brussels: European Commission, 2025. https://ec.europa.eu/newsroom/dae/redirection/document/111777.

European Commission. "DSA Transparency Database." Accessed September 10, 2025. https://transparency.dsa.ec.europa.eu/.

European Commission. "European AI Office." Accessed March 4, 2025. https://digital-strategy.ec.europa.eu/en/policies/ai-office.

European Commission. *Reporting on the application of Directive 2010/13/EU "Audiovisual Media Services Directive" as amended by Directive (EU) 2018/1808, for the period 2019-2022*. Brussels: European Commission, 2024. https://digital-strategy.ec.europa.eu/en/library/commission-report-application-audiovisual-media-services-directive.

European Commission. "The Code of conduct on countering illegal hate speech online +." Published January 20, 2025. https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online.

European Commission. "The Code of Conduct on Disinformation." Accessed February 13, 2025. https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation.

European Commission. "The Digital Services Act: Ensuring a safe and accountable online environment." Accessed February 15, 2025. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

European Commission. *REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT*

AND THE COUNCIL on the implementation of Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. Luxembourg: Publications Office of the EU, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2024%3A64%3AFIN.

European External Action Service. *EU Guidelines on Freedom of Expression Online and Offline*. Brussels: Council of the EU, 2018. https://www.eeas.europa.eu/sites/default/files/09_hr_guidelines_expression_en.pdf.

Europol. *2021 EU Internet Referral Unit Transparency Report*. Luxembourg: Publications Office of the EU, 2022. https://www.europol.europa.eu/cms/sites/default/files/documents/EU_IRU_Transparency_Report_2021.pdf.

Europol. *2023 EU Internet Referral Unit Transparency Report*. Luxembourg: Publications Office of the EU, 2025. https://www.europol.europa.eu/cms/sites/default/files/documents/2023%20EU%20Internet%20Referral%20Unit%20Transparency%20Report.pdf.

Europol. *European Union Terrorism Situation and Trend Report 2024 (EU TE-SAT)*. Luxembourg: Publications Office of the EU, 2025. https://www.europol.europa.eu/publication-events/main-reports/european-union-terrorism-situation-and-trend-report-2024-eu-te-sat.

Europol. "Europol and Telegram take on terrorist propaganda online." Published November 25, 2019. https://www.europol.europa.eu/media-press/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online.

Europol Innovation Lab. *AI and policing: The benefits and challenges of artificial intelligence For law enforcement*. Luxembourg: Publications Office of the EU, 2024. https://www.europol.europa.eu/cms/sites/default/files/documents/AI-and-policing.pdf.

Farrand, Benjamin. "How do we understand online harms? The impact of conceptual divides On regulatory divergence between the Online Safety Act and Digital Services Act." *Journal of Media Law* 16, no. 2 (2024). https://doi.org/10.1080/17577632.2024.2357463.

FATF. *Crowdfunding for Terrorism Financing*. Paris: FATF, 2023. https://www.fatf-gafi.org/content/dam/fatf-gafi/reports/Crowdfunding-Terrorism-Financing.pdf.coredownload.inline.pdf.

Fielitz, Maik, and Reem Ahmed. "It's not funny anymore. Far-right extremists' use of humour." *Radicalisation Awareness Network* (2021). https://home-affairs.ec.europa.eu/system/files/2021-03/ran_ad-hoc_pap_fre_humor_20210215_en.pdf.

Fisher, Ali. "Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence. *Perspectives on Terrorism*, 9 no. 3 (2015). https://pt.icct.nl/sites/default/files/import/pdf/2-swarmcast-how-jihadist-networks-maintain-a-persistent-online-presence-by-ali-fisher.pdf.

Frosio, Giancarlo. "Algorithmic Enforcement Tools: Governing Opacity with Due Process." In *Driving Forensic Innovation in the 21st Century*. Cham: Springer, 2024. https://doi.org/10.1007/978-3-031-56556-4_9.

Gagandeep. "Playing with Hate: How Far-Right Extremists Use Minecraft to Gamify Radicalisation." *GNET Insights* (2025). https://gnet-research.org/2025/07/02/playing- with-hate-how-far-right-extremists-use-minecraft-to-gamify-radicalisation/.

Galli, Federico, Andrea Loreggia, and Giovanni Sartor. "The Regulation of Content Moderation."

In *The Legal Challenges of the Fourth Industrial Revolution*. Cham: Springer, 2023. https://doi.org/10.1007/978-3-031-40516-7_5.

Gherbaoui, Tarik, and Martin Scheinin. "A Dual Challenge to Human Rights Law: Online Terrorist Content and Governmental Orders to Remove it." *Journal européen des droits de l'homme - European Journal of Human Rights* (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4247120.

GIFCT. "Membership." Accessed February 3, 2025. https://gifct.org/membership/.

Gillespie, Tarleton. "Reduction / Borderline content / Shadowbanning." *Yale-Wikimedia Initiative on Intermediaries & Information* (2022). https://law.yale.edu/sites/default/files/area/center/isp/documents/reduction_ispessayseries_jul2022.pdf.

Global Project Against Hate and Extremism. "Global Extremist Symbols Database." Accessed September 10, 2025. https://globalextremism.org/global-extremist-symbols-database/.

Goluvana, Valentina, and Sarah Tas. "Guardians of Digital Rights: Exploring Strategic Litigation On Data Protection and Content Moderation in the EU." *Nordic Journal of European Law* 7, no. 4 (2024). https://journals.lub.lu.se/njel/article/view/27306/24043.

Greater Manchester Police. "Young man from Liverpool convicted of preparing for acts of terrorism." Published February 23, 2024. https://www.gmp.police.uk/news/greater-manchester/news/news/2024/february/young-man-from-liverpool-convicted-of-preparing-for-acts-of-terrorism/.

Greenberg, Karen. "Counter-Radicalization via the Internet." *The Annals of the American Academy of Political and Social Science* 668, no. 1 (2016). https://doi.org/10.1177/0002716216672635.

Grippo, Valentina. *Regulating content moderation on social media to safeguard freedom of expression*. Strasbourg: Parliamentary Assembly of the Council of Europe, 2024. https://rm.coe.int/as-cult-regulating-content-moderation-on-social-media-to-safeguard-fre/1680b2b162.

Hall, Jonathan. "Rights and Values in Counter-Terrorism Online." *Studies in Conflict & Terrorism* (2023). https://www.tandfonline.com/doi/full/10.1080/1057610X.2023.2222889.

Hartgers, Menso, and Eviane Leidig. "Fighting extremism in gaming platforms: a set of design principles to develop comprehensive P/CVE strategies." *ICCT* (June 2023). https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it.

Hassan, Ghayda, Sébastien Brouillette-Alarie, Séraphin Alava et al. "Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence." *International Journal of Developmental Science* 12, no. 1-2 (2018). https://doi.org/10.3233/DEV-170233.

Held, Virginia. "The Ethics of Care." In *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press, 2009.

Herath, Chamin, and Joe Whittaker, "Online Radicalisation: Moving Beyond a Simple Dichotomy." *Terrorism and Political Violence* 35, no. 5 (2021). https://doi.org/10.1080/09546553.2021.1998008.

Holznagel, Daniel. "Follow Me to Unregulated Waters! Are Major Online Platforms Violating the DSA's Rules on Notice and Action?" Published May 30, 2024. https://doi.org/10.59704/80267b8bd7a278a4.

Huang, Justin, Jangwon Choi, and Yuqin Wan. "Politically biased moderation drives echo Chamber formation: An analysis of user-driven content removals on Reddit," *SSRN* (2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4990476.

Hunter, Sam, Alexis d'Amato, Joel Elson, Austin Doctor, and Averie Linnell. "The Metaverse as a Future Threat Landscape: An Interdisciplinary Perspective." *Perspectives on Terrorism* 18, no. 2 (June 2024). https://pt.icct.nl/sites/default/files/2024-06/Research%20article_Hunter.pdf.

Hutchinson, Jade, Amarnath Amarasingam, Ryan Scrivens, and Brian Ballsun-Stanton. "Mobilizing extremism online: comparing Australian and Canadian right-wing extremist groups on Facebook." *Behavioral Sciences of Terrorism and Political Aggression* 15, no. 2 (2023). https://doi.org/10.1080/19434472.2021.1903064.

Instagram. "How Instagram uses artificial intelligence to moderate content." Accessed August 12, 2025. https://help.instagram.com/423837189385631/?helpref=related_articles.

International Centre for Digital Threat Assessment. *Interpreting and Translating Emojis*. Surrey: Safer Schools Together, 2025. https://resources.saferschoolstogether.com/view/294238754/.

Jørgensen, Rikke Frank. "When private actors govern human rights." In *Research Handbook on Human Rights and Digital* Technology. Cheltenham: Edward Elgar Publishing, 2019.

Juszczak, Adam, and Elisa Sason. "Recalibrating Data Retention in the EU." *Eucrim*, no. 4 (2021). https://eucrim.eu/articles/recalibrating-data-retention-in-the-eu/.

Kalpakis, George, Caterina Paternoster, Marina Mancuso et al. "AI-Based Framework for Supporting Micro and Small Hosting Service Providers on the Report and Removal of Online Terrorist Content." In *Paradigms on Technology Development for Security Practitioners*. Cham: Springer, 2025. https://link.springer.com/book/10.1007/978-3-031-62083-6.

Katsaros, Matthew, Jisu Kim, and Tom Tyler. "Online Content Moderation: Does Justice Need a Human Face?" *International Journal of Human–Computer Interaction* 40, no. 1 (2024). https://doi.org/10.1080/10447318.2023.2210879.

Kaye, David. *Promotion and protection of the right to freedom of opinion and expression*. New York: UN, 2019. https://docs.un.org/en/A/74/486.

Keane, David. "Cartoon Violence and Freedom of Expression", Human Rights Quarterly 30 (2008). https://www.jstor.org/stable/20486714.

Keatinge, Tom, and Florence Keen. "Social Media and (Counter) Terrorist Finance: A Fund Raising and Disruption Tool." *Studies in Conflict & Terrorism* 42, no. 1-2 (2019). https://doi.org/10.1080/1057610X.2018.1513698.

Kerr, Dara. "TikTok to replace trust and safety team in Germany with AI and outsourced labor."

The Guardian. Published August 10, 2025. https://www.theguardian.com/technology/2025/aug/10/tiktok-trust-safety-team-moderators-ai.

Kira, Beatriz. "Regulatory intermediaries in content moderation." *Internet Policy Review* 14, no. 1 (2025). https://policyreview.info/articles/analysis/regulatory-intermediaries-content-moderation.

Klonick, Kate. "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression." *The Yale Law Journal* 129, no. 8 (June 2020). https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3639234.

Koehler, Daniel. "The Halle, Germany, Synagogue Attack and the Evolution of the Far-Right Terror Threat." *CTC Sentinel* 12, no. 11 (December 2019). https://ctc.westpoint.edu/halle-germany-synagogue-attack-evolution-far-right-terror-threat/.

Lakomy, Miron. "Let's Play a Video Game: Jihadi Propaganda in the World of Electronic Entertainment." *Studies in Conflict & Terrorism* 42, no. 4 (2019). https://doi-org.ezproxy.leidenuniv.nl/10.1080/1057610X.2017.1385903.

Lamphere-Englund, Galen, and Menso Hartgers, *CTRL+ALT+COLLABORATE: Public-Private Partnerships to Prevent Extremism in Gaming* (Luxembourg: Publications Office of the EU, 2024). https://home-affairs.ec.europa.eu/document/download/e446f013-34e1-4f74-bce6-90f661937ce9_en.

Liang, Christina Schori. "Cyber Jihad: Understanding and Countering Islamic State Propaganda." *Geneva Centre for Security Policy* (February 2015). https://www.gcsp.ch/sites/default/files/2024-12/PP2-2015%20-%20LIANG%20-%20Cyber%20Jihad%20-%20Draft%20F.pdf.

Lin, Szu-Yin, Shih-Yi Chien, Yi-Zhen Chen, and Yu-Hang Chien. "Combating Online Malicious Behavior: Integrating Machine Learning and Deep Learning Methods for Harmful News and Toxic Comments." *Information Systems Frontiers* (2024). https://doi.org/10.1007/s10796-024-10540-8.

Maaß, Sabrina, Jil Wortelker, and Armin Rott. "Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook." *Telecommunications Policy* 48, no. 5 (June 2024). https://doi.org/10.1016/j.telpol.2024.102719.

Macdonald, Stuart, Ashley Mattheis, and David Wells, *Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online*. London: Tech Against Terrorism, 2024. https://tate.techagainstterrorism.org/news/tcoaireport.

Macdonald, Stuart, and Katy Vaughan. "Moderating borderline content while respecting fundamental values." *ECTC Advisory Network Conference* (2023). https://www.europol.europa.eu/cms/sites/default/files/documents/macdonald_vaughan.pdf.

Marsoof, Althaf, Andrés Luco, Harry Tan, and Shafiq Joty. "Content-filtering AI systems limitations, challenges and regulatory approaches." *Information & Communications Technology Law* 32, no. 1 (2023). https://doi.org/10.1080/13600834.2022.2078395.

Mathur, Priyank, Clara Broekaert, and Colin Clarke. "The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence." *ICCT* (May 2024). https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence.

Mchangama, Jacob, Abby Fanlo, and Natalie Alkiviadou. *Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies*. Nashville: The Future of Free Speech, 2023. https://futurefreespeech.org/wp-content/uploads/2023/07/Community-Guidelines-Report_Latest-Version_Formated-002.pdf.

Mchangama, Jacob, and Joelle Fiss. "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship." *Justitia* (November 2019). https://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf.

Mchangama, Jacob, and Natalie Alkiviadou. "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two." *Justitia* (September 2020). https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf.

Mehra, Tanya. "Divided by Hate: Confronting Antisemitism and Islamophobia in the Netherlands." *Lawfare* (2025). https://www.lawfaremedia.org/article/divided-by-hate--confronting-antisemitism-and-islamophobia-in-the-netherlands.

Mehran, Weeda, Stephen Herron, Ben Miller, Anthony Lemieux, and Maura Conway. "Two Sides of the Same Coin? A Largescale Comparative Analysis of Extreme Right and Jihadi Online Text(s)." *Studies in Conflict & Terrorism* (2022). https://doi.org/10.1080/1057610X.2022.2071712.

Meta. "Content borderline to the Community Standards." Accessed April 30, 2025. https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards/.

Meta. "Content likely violating our Community Standards." Accessed September 10, 2025. https://transparency.meta.com/fr-fr/features/approach-to-ranking/content-distribution-guidelines/content-likely-violating-our-community-standards#.

Meta. "Hateful Conduct." Accessed September 10, 2025. https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/.

Meta. *Regulation (EU) 2022/2065 Digital Services Act (DSA): Systemic Risk Assessment and Mitigation Report for Instagram*. Menlo Park: Meta, 2024. https://transparency.meta.com/sr/dsa-sra_results_report-2024-instagram.

Meta. *Oversight Board Charter*. Menlo Park: Meta, 2019. https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.

Meta Oversight Board. *Content Moderation in a New Era for AI and Automation*. London: Meta Oversight Board, 2024. https://www.oversightboard.com/wp-content/uploads/2024/09/Oversight-Board-Content-Moderation-in-a-New-Era-for-AI-and-Automation-September-2024.pdf.

Molas, Bàrbara. "Alt-solutism: Intersections between Alt-Right Memes and Monarchism on Reddit." *ICCT* (February 2023). https://www.icct.nl/publication/alt-solutism-intersections-between-alt-right-memes-and-monarchism-reddit.

Molas, Bàrbara. "Dutch Flags and Maple Leaves: How Conspiracy Theories Created a

Transnational Far-Right." *ICCT* (July 2024). https://icct.nl/publication/dutch-flags-and-maple-leaves-how-conspiracy-theories-created-transnational-far-right.

Molas, Bàrbara, Anne Craanen, Sabrina Tripodi, Kacper Rękawek, and Thomas Renard. "Anti-Government Threats and their Transnational Connections." *ICCT* (March 2024). https://www.icct.nl/sites/default/files/2024-06/FFO-Public%20Report%20English_Final%201.pdf.

Molas, Bàrbara, and Heron Lopes. "Say it's only fictional: How the far-right is jailbreaking AI online and what can be done about it". *ICCT* (July 2024). https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it.

Mølmen, Guri, and Jacob Ravndal. "Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists." *Behavioral Sciences of Terrorism and Political Aggression* 15, no. 4 (2023). https://doi.org/10.1080/19434472.2021.1993302.

Montasari, Reza. "Machine Learning and Deep Learning Techniques in Countering Cyberterrorism." In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial* Revolution. Cham: Springer, 2024. https://doi.org/10.1007/978-3-031-50454-9_8.

Mooij, Annelieke. *Regulating the Metaverse Economy: How to Prevent Money Laundering and the Financing of Terrorism*. Cham: Springer, 2024.

Narayanan, Devesh, Mahak Nagpal, Jack McGuire, Shane Schweitzer, and David De Cremer. "Fairness Perceptions of Artificial Intelligence: A Review and Path Forward." *International Journal of Human–Computer Interaction* 40, no. 1 (2024). https://doi.org/10.1080/10447318.2023.2210890.

National Ethics Council for Social and Behavioural Sciences. *Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants*. The Netherlands: National Ethics Council for Social and Behavioural Sciences, 2018. https://nethics.nl/onewebmedia/CODE%20OF%20ETHICS%20FOR%20RESEARCH%20IN%20THE%20SOCIAL%20AND%20BEHAVIOURAL%20SCIENCES%20v2%20230518-2018.pdf.

Nationale Politie. "Aanhoudingen voor 'choking challenge'." Accessed September 10, 2025. https://www.politie.nl/nieuws/2025/januari/28/07-aanhoudingen-voor-choking-challenge.html.

NCTV. "Desinformatie" Accessed September 10, 2025. https://www.nctv.nl/onderwerpen/desinformatie.

NCTV. "Dreigingsbeeld Terrorisme Nederland." Published June 17, 2025. https://www.nctv.nl/onderwerpen/dtn.

NCTV. *Dreigingsbeeld Terrorisme Nederland: december* 2024. The Hague: Ministry of Justice and Security, 2024.

NCTV. *Memes als online wapen: Fenomeenanalyse naar het gebruik van memes door extreemrechts*. The Hague: Ministry of Justice and Security, 2024. https://www.nctv.nl/documenten/publicaties/2024/05/21/fenomeenanalyse-memes-als-online-wapen.

Openbaar Ministerie. "14 aanhoudingen in verband met opruiing tot terrorisme op sociale media." https://www.om.nl/actueel/nieuws/2025/04/22/14-aanhoudingen-in-verband-met-

opruiing-tot-terrorisme-op-sociale-media.

Overeem, Patrick. “Public service ethics and integrity: Some general lessons.” In *Ethics of counterterrorism.* Meppel: Boom, 2017.

Pauwels, Annelies, and Maarten van Alstein. “Polarisation: A short introduction.” *RAN Spotlight* (April 2022). https://home-affairs.ec.europa.eu/document/download/edf83900-6f86-4ef3-8bce-2d20e2e1d06f_en?filename=ran_spotlight_polarisation_en.pdf.

Pauwels, Annelies, and Merlina Herbach. “Buy It, Steal It, Print It: How Right-Wing Extremists In Europe Acquire Firearms And What To Do About It.” *ICCT* (December 2024). https://icct.nl/publication/buy-it-steal-it-print-it-how-right-wing-extremists-europe-acquire-firearms-and-what-do.

Pauwels, Annelies, and Will Baldet. “The Role of Hotbeds of Radicalisation.” *RAN Spotlight* (April 2022). https://home-affairs.ec.europa.eu/document/download/edf83900-6f86-4ef3-8bce-2d20e2e1d06f_en?filename=ran_spotlight_polarisation_en.pdf.

Peeters, Timo, Ron van Wonderen, Dominique Burggraaff, and Nienke de Wit. *Online extreemrechtse radicalisering: Handvatten voor een preventieve aanpak*. Utrecht: Verwey Jonker Instituut, 2022. https://www.verwey-jonker.nl/wp-content/uploads/2022/12/121650_Online-extreemrechtse-radicalisering.pdf.

Pírková, Eliška, Marlena Wisniak, and Karolina Iwańska. *Towards Meaningful Fundamental Rights Impact Assessments Under The DSA*. The Hague & Brussels: European Center for Not-for-Profit Law & Access Now Europe, 2023. https://ecnl.org/sites/default/files/2023-09/Towards%20Meaningful%20FRIAs%20under%20the%20DSA_ECNL%20Access%20Now.pdf.

Popović, Dušan. “The Digital Platforms’ Sisyphean Task: Reconciling Content Moderation and Freedom of Expression.” In *Repositioning Platforms in Digital Market Law*. Cham: Springer, 2024. https://doi.org/10.1007/978-3-031-69678-7_4.

Prem, Erich, and Brigitte Krenn. “On Algorithmic Content Moderation.” In *Introduction to Digital Humanism*. Cham: Springer, 2023. https://doi.org/10.1007/978-3-031-45304-5_30.

Rai, Steven. “Beyond the Collective: Understanding Terrorgram’s efforts to infiltrate the mainstream on Telegram.” Published August 24, 2024. https://www.isdglobal.org/digital_dispatches/beyond-the-collective-understanding-terrorgrams-efforts-to-infiltrate-the-mainstream-on-telegram/.

Clarke, Colin, Camden Carmichael, and Seamus Hughes. “Why the Terrorgram Collective Designation Matters.” *Lawfare* (2025). https://www.lawfaremedia.org/article/why-the-terrorgram-collective-designation-matters.

Reddit. “Content Moderation, Enforcement, and Appeals.” Accessed August 12, 2025. https://support.reddithelp.com/hc/en-us/articles/23511059871252-Content-Moderation-Enforcement-and-Appeals.

Reddit. “How does Reddit fight the dissemination of terrorist content?” Accessed September 10, 2025. https://support.reddithelp.com/hc/en-us/articles/19003525756564-how-does-Reddit-fight-the-dissemination-of-terrorist-content.

Rijksoverheid. “Ministerraad stemt in met wetsvoorstel om verheerlijken van terrorisme strafbaar te stellen.” Accessed September 19, 2025. https://www.rijksoverheid.nl/actueel/nieuws/2025/06/20/ministerraad-stemt-in-met-wetsvoorstel-om-verheerlijken-van-terrorisme-strafbaar-te-stellen.

Rojszczak, Marcin. “Gone in 60 Minutes: Distribution of Terrorist Content and Free Speech in the European Union.” *Democracy and Security* 20, no. 2 (2023). https://www.tandfonline.com/doi/abs/10.1080/17419166.2023.2250731.

Romero-Moreno, Felipe. “Generative AI and deepfakes: a human rights approach to tackling harmful content.” *International Review of Law, Computers & Technology* 38, no. 3 (2024). https://doi.org/10.1080/13600869.2024.2324540.

Rottier, Melissa. *Onderzoek naar de verificatie van kinderpornografisch en terroristisch materiaal ten behoeve van databases*. Rotterdam: ATKM, 2025. https://www.atkm.nl/documenten/2025/07/23/onderzoek-naar-kwaliteit-databases-die-moeten-voorkomen-dat-schadelijk-materiaal-wordt-gedeeld.

Saltman, Erin, and Micalie Hunt. “Borderline Content: Understanding the Gray Zone.” *GIFTC* (2023). https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf.

Schindler, Hans-Jakob. “Emerging challenges for combating the financing of terrorism in the European Union: financing of violent right-wing extremism and misuse of new technologies.” *Global Affairs* 7, no. 5 (2021). https://doi.org/10.1080/23340460.2021.1977161.

Schneider, Philipp, and Marian-Andrei Rizoiu. “The effectiveness of moderating harmful online content.” *PNAS* 120, no. 34 (2023). https://www.pnas.org/doi/10.1073/pnas.2307360120.

Schulz, Wolfgang, Karmen Turk, Bertrand de la Chapelle, et al. *Algorithms and human rights – Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. Strasbourg: Council of Europe, 2018. https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html.

Siegel, Alexandra. “Online Hate Speech.” In *Social Media and Democracy: The State of the Field and Prospects for Reform*. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108890960

Siegel, Daniel. “AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State’s Generative AI Digital Arsenal.” *GNET Insights* (2024). https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/.

Sobol, Ilya. “Glorification of Terrorist Violence at the European Court of Human Rights.” *Human Rights Law Review* 24, no. 3 (2024). https://academic.oup.com/hrlr/article/24/3/ngae017/7696469.

Sombatpoonsiri, Janjira, and Sangeeta Mahapatra. “Regulation or Repression? Government Influence on Political Content Moderation in India and Thailand.” *Carnegie Endowment for International Peace* (July 2024). https://carnegieendowment.org/research/2024/07/india-thailand-social-media-moderation?lang=en.

Stoeldraaijers, Charlie, Elanie Rodermond, Fabienne Thijs, Rutger Leukfeldt, and Frank Weerman, *Radicale reclame op sociale media: Een onderzoek naar online rekrutering door en voor extremistische groepen*. Amsterdam: Nederlands Studiecentrum Criminaliteit en Rechtshandhaving, 2024. https://repository.wodc.nl/bitstream/handle/20.500.12832/3382/3417-radicale-reclame-op-sociale-media-volledige-tekst.pdf?sequence=1&isAllowed=y

Tech Against Terrorism. "Patterns of Terrorist Online Exploitation." *TCAP Insights* (April 2023). https://26492205.fs1.hubspotusercontent-eu1.net/hubfs/26492205/260423%20TCAP%20INSIGHTS%20-%20FINAL.pdf.

Thakkar, Mona, and Anne Speckhard. "Caliphate AI - IS/ISKP Supporters Harness Generative AI for Propaganda Dissemination." *International Center for the Study of Violent Extremism* (July 2024). https://www.researchgate.net/publication/382025204_Caliphate_AI_-ISISKP_Supporters_Harness_Generative_AI_for_Propaganda_Dissemination.

The Future of Free Speech. *Preventing "Torrents of Hate" or Stifling Free Expression Online? An Assessment of Social Media Content Removal in France, Germany, and Sweden*. Nashville: The Future of Free Speech, 2024.

TikTok. "Content Moderation." Accessed August 12, 2025. https://www.tiktok.com/euonlinesafety/en/content-moderation/.

TikTok. *DSA Risk Assessment Report 2023*. Dublin: TikTok, 2023. https://panoptykon.org/sites/default/files/2025-01/tiktok-dsa-risk-assessment-report-2023.pdf.

TikTok. "Safety and Civility," Accessed September 15, 2025. https://www.tiktok.com/community-guidelines/en-GB/safety-civility?cgversion=2025H2update.

TikTok. "Terms of Service." Accessed September 15, 2025. https://www.tiktok.com/legal/page/eea/terms-of-service/en.

TSPA. "Enforcement Methods and Actions." Accessed January 16, 2025. https://www.tspa.org/curriculum/ts-fundamentals/policy/enforcement-methods/.

Tuchtfeld, Erik. "Case law on content moderation and freedom of expression." *The Global Freedom of Expression Special Collection of the Case Law on Freedom of Expression* (June 2023). https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2023/06/GFoE_Content-Moderation.pdf.

Turillazzi, Aina, Mariarosaria Taddeoa, Luciano Floridi, and Federico Casolari. "The Digital Services Act: An Analysis of Its Ethical, Legal, and Social Implications." *Law, Innovation and Technology* 15, no. 1 (2023). https://www.tandfonline.com/doi/full/10.1080/17579961.2023.2184136?src=recsys.

Types Digital. "AI Takes Over Content Moderation at TikTok." *Medium*. Published October 17, 2024. https://medium.com/@types24digital/ai-takes-over-content-moderation-at-tiktok-d4131bc6ae74.

UK Government. "New definition of extremism (2024)." Published March 14, 2024. https://www.gov.uk/government/publications/new-definition-of-extremism-2024/new-definition-of-extremism-2024.

UN Human Rights Council. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*. New York & Geneva: UN, 2011. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

UN Office on Genocide Prevention and the Responsibility to Protect. *United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences*. New York: UN, 2020. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

Urman, Aleksandra, and Stefan Katz. "What they do in the shadows: examining the far-right networks on Telegram." *Information, Communication & Society* 25, no. 7 (2022). https://doi.org/10.1080/1369118X.2020.1803946.

Van den Branden, Britt, Sophie Davidse, and Eva Smit. "In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression (Part 1 of 3)." Published August 2, 2021. https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/#:~:text=In%20Article%202(g)%20DSA,tools%20used%20in%20content%20moderation.

Van Ginkel, Bibi. "Responding to Cyber Jihad: Towards an Effective Counter Narrative." *ICCT* (March 2015). https://www.icct.nl/sites/default/files/2022-12/ICCT-van-Ginkel-Responding-To-Cyber-Jihad-Towards-An-Effective-Counte.pdf.

Van Ginkel, Bibi, Rik Scheele, Rebecca Visser, and Menso Hartgers. "Online Monitoring of Radicalisation and (Violent) Extremism: Mapping Legal and Policy Challenges for Online P/CVE Work." *RAN Conclusion Paper* (2023).

Van Huijstee, Mariëtte, Wouter Nieuwenhuizen, Mathilde Sanders, Eef Masson, and Pieter van Boheemen. *Online ontspoord – Een verkenning van schadelijk en immoreel gedrag op het internet in Nederland*. The Hague: Rathenau Instituut, 2021.

Van Wonderen, Ron, Dominique Burggraaff, Soenita Ganpat, Gijs van Beek, and Olivier Cauberghs. *Rechtsextremisme op sociale mediaplatforms? Ontwikkelingspaden en handelingsperspectieven*. Utrecht: Verwey Jonker Instituut, 2023. https://repository.wodc.nl/bitstream/handle/20.500.12832/3304/3341-rechtsextremisme-op-sociale-media-platforms-volledige-tekst.pdf?sequence=1&isAllowed=y.

Van Wonderen, Ron, Joline Verloove, and Hanneke Felten. *Theorieën en aanpakken van polarisatie — Verkorte rapportage van bevindingen en conclusies*. Utrecht: Kennisplatform Integratie & Samenleving, 2020. https://www.kis.nl/sites/default/files/2022-06/theorieen-en-aanpakken-van-polarisatie-samenvatting.pdf.

Visser, Rebecca. "Crowdfunding Conspiracists: Grassroots Giving to January 6 Participants." *ICCT* (December 2024). https://icct.nl/publication/crowdfunding-conspiracists-grassroots-giving-january-6-participants.

Völker, Teresa. "How terrorist attacks distort public debates: a comparative study of right-wing

and Islamist extremism." *Journal of European Public Policy* 31, no. 11, (2023). https://doi.org/10.1080/13501763.2023.2269194.

Walther, Samantha, and Andrew McCoy. "US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism." *Perspectives on Terrorism* 15, no. 2 (April 2021). https://www.icct.nl/sites/default/files/import/publication/walther-and-mccoy-.pdf.

Watkin, Amy-Louise. "Developing a Responsive Regulatory Approach to Online Terrorist Content on Tech Platforms." *Studies in Conflict & Terrorism* (2023). https://doi.org/10.1080/1057610X.2023.2222891.

Weimann, Gabriel, Alexander Pack, and Gal Rapaport. "White Jihad: Fused Extremism?" *Terrorism and Political Violence* (2025). https://gnet-research.org/2024/01/19/764-the-intersection-of-terrorism-violent-extremism-and-child-sexual-exploitation/.

Westerstrand, Salla. "Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence." *Science and Engineering Ethics 30, no. 46 (2024).* https://link.springer.com/article/10.1007/s11948-024-00507-y?trk=public_post_comment-text.

Williams, Heather, Alexandra Evans, Jamie Ryan, Erik Mueller, and Bryce Downing. "The Online Extremist Ecosystem: Its Evolution and a Framework for Separating Extreme from Mainstream." *RAND Cooperation* (December 2021). https://www.rand.org/pubs/perspectives/PEA1458-1.html.

Won, Ye Bin, and Jonathan Lewis. "Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts." *GNET Insights* (2021). https://gnet-research.org/2021/04/06/male-supremacism-borderline-content-and-gaps-in-existing-moderation-efforts/.

Yao, Deborah. "EU AI Act Would Scrutinize Many 'General' AI Models – SXSW 2024." *AI Business. Published March 13, 2024.* https://aibusiness.com/responsible-ai/eu-ai-act-would-scrutinize-many-general-ai-models-sxsw-2024#close-modal.

Annex 1: List of Interviewees

Position	Organisation
Prosecutor	Public Prosecutor's Office
President of the Board	ATKM
Expert	EU Internet Referral Unit
Expert	GIFCT
Expert	GIFCT
Big Tech platform researcher	NGO in the field of privacy issue
Expert	NGO in the field of privacy issues
Expert	Textgain/EOOH
Expert on human rights	NGO in the field of human rights

Annex 2: List of Interview Questions

Feasibility Study: Development of an Assessment Framework for Terrorist and Borderline Content Online

In the June 2024 Terrorism Threat Assessment Netherlands (DTN), the Dutch National Coordinator for Counterterrorism and Security (NCTV) warned that extremist ideology and terrorist content are being spread via social media and other online platforms, and that these platforms enable communication between extremists and terrorists, recruitment, and direction of attackers, as well as the planning of attacks.

An assessment framework for terrorist and borderline online content can contribute to a better understanding and recognition of this harmful online content, making detection easier. Hence, the NCTV, through the WODC Research and Data Centre of the Dutch Ministry of Justice, commissioned the International Centre for Counter-Terrorism (ICCT) to conduct a feasibility study on the development of an assessment framework for terrorist and borderline content online.

A challenge related to the detection of borderline online content is the oftentimes implicit character of the message, making it not always easy and immediately detectable. Such harmful content is also known as borderline content. The research aims to examine whether tools and methods exist or can be developed to better detect and identify both implicit and explicit harmful online content, without compromising the fundamental right to freedom of expression. The research aims to contribute to the development of more effective strategies that help increase the resilience of online platforms and society against terrorist and extremist content, thus countering online radicalisation. The development of an assessment framework should ultimately enable online platforms to carry out content moderation without compromising the fundamental right to freedom of expression. If the development of an interpretation framework according to the above specifications is deemed feasible, it will require additional research. Thus, the actual development of the interpretation framework is not part of this project.

Questionnaire

<u>General questions for all interviewees:</u>

1. Please describe your job briefly?
2. Please explain how terrorist and extremist online content is relevant to your daily work.
3. What do you consider to be borderline content? And do you consider it a concern? Why and to whom?

<u>Questions for different stakeholders:</u>

Online Service Providers

4. How do you define illegal, terrorist, extremist and borderline content? And are these definitions included in your Terms of Service?
5. How do you use these definitions in practice?
6. How do you detect terrorist and borderline content? (Manual/AI) What are the main benefits, challenges, and gaps of your policy?
7. Do you perceive any difference in detecting terrorist and borderline content?
8. Are you able to detect both explicit and implicit borderline content (such as humor or irony?
9. Do you use different methods to detect explicit and implicit borderline content?
10. Is the mechanism to detect terrorist or borderline content equally applicable to different

forms of terrorism (jihadi, far-right, far left, anti-institutionalism)? Do you observe any differences in detection based on the underlying ideologies?
11. How often do you update your methods of detection? How do you train both AI and humans to stay up to date with new forms, memes, ideology etc?
12. How do you calibrate your detection method?
13. How reliable is the method you use to detect borderline content? What approx. % is false negatives or positives? How do you correct false negatives or positives?
14. How do you ensure that freedom of expression is protected while assessing borderline content?
15. What kind of content moderation measures (i.e. removal, account suspension, lower visibility) can you take? And how frequently do you use these different measures?
16. How do you inform the users of content moderation measures you take, and what procedures do you have in place to challenge these decisions?
17. What kind of reporting mechanisms are available on your platform?
18. When and under which conditions would you report the content to the police? Do you have outside of case by case reporting, meetings with state authorities? How often do these meetings take place?
19. In the past year, how often have you reported specific online content to the police?
20. Do you have meetings with experts, CSOs and other platforms to discuss the issue of borderline content? How often do these meetings take place?
21. Has the TCO Regulation facilitated your work in detecting terrorist content (either by guidance on defining the scope or because of the take down notices)?
22. Has the DSA facilitated your work in detecting other illegal content?
23. The DSA also refers to systemic risks and the need to moderate content that falls under that category. Borderline content could qualify as such. How are you dealing with that, and/or detecting and moderating such content?
24. How have the DSA or the TCO regulation affected your content moderation and reporting?
25. How much borderline content do you already detect and remove before receiving removal requests?

National Authorities

26. How is the cooperation between national authorities and online service providers around the detection and moderation of borderline content?
27. How reliable do you consider the detection mechanisms used by online service providers in relation to terrorist and borderline content? Is there a difference between the two forms of content? And what is the difference between online service providers in how they deal with this?
28. Do you think that online service providers are equipped to detect explicit and implicit borderline content?
29. Do you think that online service providers are equipped to detect borderline content related to all different ideologies (i.e. RWX, LXW, Islamist, anti-institutional)?
30. Are the online service providers transparent enough about their detection methods and the moderation measures that they take?
31. Do you think that online service providers sufficiently protect the freedom of expression and other fundamental rights of users when implementing their detection and removal mechanisms with respect to borderline content?
32. How have the TCO and DSA affected the detection and removal of borderline content from the perspective of national authorities?

Civil Society Organisations

33. Do you have a structural dialogue with online service providers? On what topics and how often?

34. Do you believe that CSOs can and should have a role in the detection and moderation of terrorist and borderline content? Why?
35. Can CSOs contribute to the spread of terrorist or borderline content?
36. How reliable do you consider the detection mechanisms of online service providers in relation to terrorist and borderline content? Do you perceive a difference between online service providers and types of content?
37. Do you think that online service providers are equipped to detect explicit and implicit borderline content?
38. Do you think that online service providers are equipped to detect borderline content related to all different ideologies (i.e. RWX, LXW, Islamist, anti-institutional)?
39. Are the online service providers transparent enough about their detection methods and the moderation measures that they take? If not, in which areas can they make improvements?
40. Do you think that online service providers sufficiently protect the freedom of expression and other fundamental rights of users with respect to borderline content?
41. Has the TCO Regulation improved the detection and moderation of terrorist content online?
42. Has the DSA improved the detection of other illegal content?
43. The DSA also refers to systemic risks and the need to moderate content that falls under that category. Borderline content could qualify as such. What is your assessment of this possibility regarding the detection and moderation of borderline content online?
44. Do you believe that small online service providers are sufficiently equipped and receive sufficient assistance to detect and moderate terrorist and borderline content online?
45. Do you think that current European legal frameworks such as the TCO and DSA, provide sufficient possibilities and safeguards for the use of automated content detection and moderation?

Questions to all stakeholders regarding a Codebook

46. Do you think that it would be feasible to create a codebook or systematic analytical tool that allowed internet service providers to more effectively identify borderline content?
47. Do you think that it would be desirable to create a codebook or systematic analytical tool that allowed internet service providers to more effectively identify borderline content?
48. What do you think the main benefits and challenges of such a codebook would be?
49. Who should be involved in developing such a codebook?
50. *For Online Service Providers: If such a uniform codebook werfe developed, would you implement and use it?*

Final Remarks

51. Is there anything you would want to add to or revise in your previous answers?

Annex 3: Agenda Expert Meeting

Expert Meeting – Draft Agenda

15 May 2025, 13.00 – 17.00 hrs

Pulchri Studio, Lange Voorhout 14, The Hague

13.00 Welcome and introduction

13.15 – 14.15 Session 1: Setting the scene and discussing scope and definition

Challenge 1: Vague and unclear definitions of borderline content by states and platforms.
Challenge 2: Keeping up with the continuously evolving field of the way terrorist/extremist/borderline content is portrayed.
Challenge 3: Different perceptions regarding who needs to moderate borderline content.
Challenge 4: How to distinguish between protected speech and hate speech?
Challenge 5: Identification of key elements to recognise borderline and/or harmful content, and the lack of consensus on these elements.

14.15 – 14.30 Break

14.30 – 15.30 Session 2: Discussing detection and moderation methods; Can we identify obstacles versus smart solutions?

Challenge 1: Evasive tactics used by posters to avoid detection are constantly changing.
Challenge 2: Keeping detection feasible, dealing with dilemmas such as volume versus precision; dealing with false positives versus false negatives; and choosing between automated versus manual detection.
Challenge 3: Platform in charge of detection versus users in charge of detection
Challenge 4: Moderating content while respecting freedom of expression (keeping in line with principles such as necessity and proportionality, and right to appeal)

15.30 – 15.45 Break

15.45 – 16.45 Session 3: Discussing the feasibility and desirability of an assessment framework to unequivocally detect borderline content

Challenge 1: Overcoming legal challenges to unequivocally detect borderline content
Challenge 2: Overcoming practical/technical challenges to unequivocally detect borderline content
Challenge 3: Overcoming ethical challenges to detect borderline content
Challenge 4: Connecting the appropriate moderation methods to the qualification of the kind of content.

16.45 – 17.00 Concluding remarks

Annex 4: Definitions of Hate Speech

Entity	Definition hate speech
UN Strategy and Plan of Action on Hate Speech	any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Council of Europe Recommendation CM/Rec(2022)16[1] of the Committee of Ministers to member States on combating hate speech	hate speech is understood as all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as “race”, colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation.
EU Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law	1. Each Member State shall take the measures necessary to ensure that the following intentional conduct is punishable: (a) publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin; (b) the commission of an act referred to in point (a) by public dissemination or distribution of tracts, pictures or other material; (c) publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in Articles 6, 7 and 8 of the Statute of the International Criminal Court, directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group; (d) publicly condoning, denying or grossly trivialising the crimes defined in Article 6 of the Charter of the International Military Tribunal appended to the London Agreement of 8 August 1945, directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group.
EU Code of Conduct on Countering Illegal Hate Speech Online +	Illegal hate speech as defined by applicable laws, including the Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, as transposed in national jurisdictions, as well as possible forthcoming updates to this Framework Decision, where relevant.
Cambridge Dictionary	public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation

Instagram	hateful conduct as direct attacks against people – rather than concepts or institutions – on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. Additionally, we consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks (Tier 1 below), though we do allow commentary on and criticism of immigration policies. Similarly, we provide some protections for non-protected characteristics, such as occupation, when they are referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as frequently used proxies for protected characteristics. **Hate entity** – defined as an organisation or individual that spreads and encourages hate against others based on their protected characteristics. The entity's activities are characterised by at least some of the following behaviours: Violence, threatening rhetoric or dangerous forms of harassment targeting people based on their protected characteristics; Repeated use of hate speech; Representation of hate ideologies or other designated hate entities, and/or Glorification or Support of other designated Hate Entities or Hate Ideologies.
Reddit	allow posts against marginalised or vulnerable groups which include, but are not limited to, groups based on their actual and perceived race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families.
TikTok	prohibits hate speech, hateful behaviour and the promotion of hateful ideologies, including both explicit or implicit content targeting a protected group. Hateful ideologies are systems of beliefs that exclude, oppress, or otherwise discriminate against individuals based on their protected attributes. Protected groups means individuals or communities that share protected attributes. Protected attributes mean personal characteristics that you are either born with, are immutable, or that would cause severe psychological harm if you were forced to change them or were attacked because of them. This includes: caste, ethnicity, national origin, national origin, race, religion, tribe, immigration status, gender, gender identity, sex, sexual orientation, disability and serious disease. In addition, we also provide some protections related to age, and may consider other protected attributes when we have additional context, such as specific regional information provided to us by a local non-governmental organisation (NGO). The attributes listed above are informed by the Universal Declaration of Human Rights and international conventions. Content may be ineligible for the FYF when it indirectly demeans protected groups.

NL : artikel 137d	1.Hij die in het openbaar, mondeling of bij geschrift of afbeelding, aanzet tot haat tegen of discriminatie van mensen of gewelddadig optreden tegen persoon of goed van mensen wegens hun ras, hun godsdienst of levensovertuiging, hun geslacht, hun seksuele gerichtheid of hun handicap, wordt gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie. 2 Indien het feit wordt gepleegd door een persoon die daarvan een beroep of gewoonte maakt of door twee of meer verenigde personen wordt gevangenisstraf van ten hoogste vier jaren of geldboete van de vierde categorie opgelegd.
NL : artikel 137c	1.Hij die zich in het openbaar, mondeling of bij geschrift of afbeelding, opzettelijk beledigend uitlaat over een groep mensen wegens hun ras, hun godsdienst of levensovertuiging, hun seksuele gerichtheid of hun handicap, wordt gestraft met gevangenisstraf van ten hoogste een jaar of geldboete van de derde categorie. 2.Met dezelfde straf wordt gestraft degene die zich in het openbaar, mondeling of bij geschrift of afbeelding, opzettelijk beledigend uitlaat over een groep mensen als omschreven in het eerste lid: a. door het vergoelijken van een van de feiten als omschreven in de artikelen 3 tot en met 6, 7, tweede lid, en 8 tot en met 8b van de Wet internationale misdrijven of een van de feiten als omschreven in artikel 6 van het Handvest van de Internationale Militaire Rechtbank, gehecht aan het Verdrag van Londen van 8 augustus 1945; b. door het ontkennen of verregaand bagatelliseren van een van de feiten als omschreven in de onder a genoemde artikelen, voor zover dat feit bij onherroepelijke beslissing is vastgesteld door een internationaal gerecht dat zijn rechtsmacht ontleent aan een verdrag waarbij het Koninkrijk partij is of door de Nederlandse rechter. 3.Indien het feit wordt gepleegd door een persoon die daarvan een beroep of gewoonte maakt of door twee of meer verenigde personen wordt gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie opgelegd.

Annex 5: Legitimisation for Assessing the Suitability of the Different Online Platforms

Legitimisation of online platforms for assessing the suitability for using pilot codebook		
General considerations for selecting platforms		
Purpose of this part of the research for the overall objective	The overall objective is to consider the feasibility of setting up an assessment framework of terrorist, violent extremist and borderline content, which is effective, operational and respects the freedom of speech right. The OSINT contactless research as part of this research project merely serves the purpose of piloting a selection of indicators that could assist in identifying the mentioned content. It is therefore important to select platforms for piloting the indicators that would allow testing the applicability on different ideologies and different forms of content. It is, therefore not an objective of this research project to get better insight into the level, amount, type of terrorist, extremist or borderline content on key platforms.	

Considerations related to content	This study focuses on borderline content, an area that is largely under-researched. Hence, the research should also focus on platforms where such borderline content can be found. Previous research found that when faced with content moderation on a platform, users first adapt by engaging with and creating more borderline content. In the next step, they navigate to less moderated or (partially) encrypted platforms where they can engage with more explicit terrorist/extremist content and do not need to resort to borderline framings. It would also be more beneficial to look at platforms were borderline content can still be found online. YouTube and Facebook for example were found to identify and remove/lower the visibility of a lot of content through AI, including many false positives (over-blocking) so it is less likely that borderline or explicit content is still available on these sites. A look at previous research indicates that RWEs in the Netherlands were particularly active on Facebook, YouTube, Twitter and to a certain extent also on Instagram, see Timo Peeters et al., *Online extreemrechtse radicalisering: Handvatten voor een preventieve aanpak* (Utrecht: Verwey Jonker Instituut, 2022), https://www.verwey-jonker.nl/wp-content/uploads/2022/12/121650_Online-extreemrechtse-radicalisering.pdf. A more recent study on online radicalisation and propaganda also mentions Facebook, Twitter (X), TikTok, Instagram, and YouTube as the most popular mainstream platforms for extremists to use for their purposes with Telegram and Discord being the more fringe but popular ones too, see Charlie Stoeldraaijers et al., *Radicale reclame op sociale media: Een onderzoek naar online rekrutering door en voor extremistische groepen* (Amsterdam: Nederlands Studiecentrum Criminaliteit en Rechtshandhaving, 2024), https://repository.wodc.nl/bitstream/handle/20.500.12832/3382/3417-radicale-reclame-op-sociale-media-volledige-tekst.pdf?sequence=1&isAllowed=y; On softening of tone, see Erin Saltman, and Micalie Hunt, "Borderline Content: Understanding the Gray Zone," *GIFTC* (2023): 7, https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf. On mainstream (moderated) platforms and less moderated/niche platforms used by extremists and their different behaviour on the different platforms, see Heather Williams et al., "The Online Extremist Ecosystem: Its Evolution and a Framework for Separating Extreme from Mainstream," *RAND Cooperation* (December 2021), https://www.rand.org/pubs/perspectives/PEA1458-1.html. Furthermore, we are keeping in mind that platforms that belong to the same company, for example Instagram and Facebook both belonging to Meta, might show - to a certain extent- similarities in the policies and practices of content moderation.

Budget limitations	Due to the budget limitations of this project, the scope of the OSINT research part of this project is rather limited, and is merely used as a pilot to provide an additional option to reflect on the feasibility of setting up an assessment framework. The scope is therefore limited to the assessment of content on only three platforms, and content posted within the timeframe of one week. The language is furthermore limited to Dutch and English, and the geographical focus is the Netherlands. Due to budget limitations, the research can only focus on platforms on which OSINT is easily possible without deploying additional tools that come at a cost. Concerning the geographical scope of the Netherlands, one must also take into account that the relevance of a platform in the Netherlands is not only determined by the number of Netherlands-based users, but also whether Netherlands-related terrorist/extremist/borderline content has been shared on these platforms in the past.
Uniformity/consistency in research methodology	Related to the budget limitations, the uniformity and consistency in research methodology, in other words the methods of scraping, saving, anonymising, coding and analysing, needs to be as uniform as possible for reasons of efficiency. In addition, considering the fact that the objective is to get an understanding of the way platforms moderate the content, we compare their policies with our own observations, hence, an important aspect of our overall consideration in selecting the platforms also relates to the cooperative attitudes of the platforms. Borderline content can be found across different categories of content, first and foremost is hate speech. Anticipating the chance that we will get limited interaction/response from the platform we also keep in mind that certain platforms have signed on voluntarily to the Code of Conduct on Hate Speech+. The implementation reports already provide interesting data that we can use. Companies that have signed on are Dailymotion, Facebook, **Instagram**, Jeuxvideo.com, LinkedIn, Snapchat, Rakuten Viber, **TikTok**, Twitch, X and YouTube. It would also be helpful to take into account which companies are member of GIFCT and are included in their research: **Instagram**, **TikTok**, Discord, Snapchat, Facebook, LinkedIn.
Ethical limitations	This research relies on the assumption that context is paramount in determining whether certain conduct is legally permissible or must be removed as terrorist/extremist. Hence, automised data scraping was not a preferred option. Instead, all data is being scraped manually. This comes with certain ethical implications for researchers as well as the platform users. The latter do not know that they are being studied, hence data and information relevant for the study can only be reproduced without revealing identifying information. The fact that these users are studied without their consent and awareness also restricts the extent to which the researchers can engage with them. Interaction is thus strictly prohibited. See: Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants, adopted on 23 May 2018 in the Social Sciences Discipline Consultation. We furthermore respect the do no harm principle in our research approach. This means that we do not only consider this principle when considering privacy/ reputational issues for the target audience we are researching, but we also keep the safety of our staff in mind. This means that for OSINT research we only do anonymous and contactless research. OSINT into content on platforms for which the researchers need to identify themselves, and get vetted before access is allowed, are therefore excluded.

Annex 6 : Codebook

General Information									
N/A									
Post ID	Date	Language	Platform	Format	Account type	Trigger event	Likes / votes	Comments / replies	Any additional notes

STEP 1. Terrorist Content (TC)										
Incitement to commit/participate in a terrorist offence (IT)										
Intended audience is a **group** (IT)	Audience sufficiently clear to include people with **ideological ties** (IT)	Content is spread in **public** (IT)	Advocate the commission of a terrorist offence (IT)	Awareness of the likelihood that the audience could commit a terrorist act - **capacity** (IT)	Deliberately creates the likelihood of a **serious** terrorist offence regardless of whether executed or not (IT)	Direct or indirect **causal link** between incitement and terrorist act (IT)	**Imminent** risk that a terrorist act will be committed (IT)	**Knowledge of or reference to** terrorist activities (IT)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (IT)	**Presence of IT**

STEP 1. Terrorist Content (TC)										
Glorification of a terrorist act (GT)					Recruitment for a terrorist organisation (RT)					**Presence of terrorist content (TC)**
Justification of past/ future serious terrorist offence (GT)	Deliberately creating an **atmosphere of fear** (GT)	**Knowledge of or reference to** terrorist activities (GT)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (GT)	**Presence of GT**	Directly or indirectly addressing an audience - individual or group (RT)	Call to supply information/ material resources/ funding/ human resources to a terrorist group (RT)	**Knowledge of or reference to** a specific terrorist group, one or more serious criminal acts committed by the referenced terrorist group, one or more concrete aims of the terrorist group (RT)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (RT)	**Presence of RT**	

STEP 2. Explict Illegal Content (EIC)						
Hate speech (HS)						
Content is spread in **public** (HS)	**Discriminates against a group or individual** (HS)	Discriminates **on the basis of** race, gender, religion, belief, sexual orientation, physical or mental disability (HS)	**Insults/incites** hatred or violence (HS)	**Knowledge of or reference to** discriminatory acts or speech (HS)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (HS)	**Presence of HS**

STEP 2. Explict Illegal Content (EIC)								
Incitement to violence (IV)								
Content is spread in **public** (IV)	**Direct and imminent causal link** between incitement and serious violent offence (IV)	**Advocates commission** of a serious violent offence (IV)	Awareness of likelihood that audience could commit a serious violent offence - **capacity** (IV)	Deliberately creates the likelihood that a **serious** violent offence will be comitted regardless of execution or not (IV)	Intended audience is a **group** (IV)	**Knowledge or reference** to serious acts of violence (IV)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (IV)	**Presence of IV**

STEP 2. Explict Illegal Content (EIC)						
Denial, downplaying or justification of international crimes (DIC)						**Presence of Explicit Illegal Content (EIC)**
Content is spread in **public** (DIC)	Against a **group or individual** (DIC)	**Denies, justifies, or downplays** core international crimes established irrevocably by a court (DIC)	**Knowledge or reference** to the fact that crimes constitute core international crimes (DIC)	**Not** a satire, parody, artistic expression, legitimate contribution to the public discourse, legitimate commemoration of historical events, colonial past or decolonialisation (if it is not = 1; if it is=0) (DIC)	**Presence of DIC**	

STEP 3. Implicit Extremist Content (IHC)							
Intent to conceal meaning (CM)							
Ambiguous or coded language (CM)	Altered text or misspelling (CM)	Misleading cover image for video (CM)	Blurred or altered image - e.g. marker function to hide certain words (CM)	Use of humour or irony (CM)	Other (CM)	Other (specify)	**Presence of CM**

STEP 3. Implicit Extremist Content (IHC)								
Harmful alliances/affiliations (HA)								
Hateful / extremist symbols (far right)	Hateful / extremist symbols (Islamist)	Hateful / extremist emojis (far right)	Hateful / extremist emojis (Islamist)	Hateful / extremist coded slogans, slang, or acronyms (far-right)	Hateful / extremist coded slogans, slang, or acronyms (Islamist)	Other (HA)	Other (specify)	**Presence of HA**

STEP 3. Implicit Extremist Content (IHC)									
Problematic references to historical/current context (PR)									
Denial or questioning of proven past or present crimes (PR)	False or misleading claims about past or present crimes (PR)	Justification of current or potential future crimes through references to past crimes (PR)	Falsified historical claims aimed at denying the existence or territorial legitimacy of a state (PR)	Glorification or positive portrayal of individuals or groups involved in crimes (PR)	Glorification or positive portrayal of public figures known for spreading hateful or extremist narratives (PR)	Promotion or endorsement of books, films, essays, or any other products known for spreading or supporting any of the above narratives (PR)	Other (PR)	Other (specify)	**Presence of PR**

STEP 3. Implicit Extremist Content (IHC)								
Implicit action triggers (AT)								
Instilling a perceived need for retaliation (AT)	Instilling a perceived need for self-defence (AT)	Instilling a perceived need to protect a group from reputational damage (AT)	Instilling the idea of compensation or reward for undertaking action (AT)	Exerting peer pressure/respect for a code of honour (AT)	Providing sacred justification for action (AT)	Other (AT)	Other (specify)	**Presence of AT**

STEP 3. Implicit Extremist Content (IHC)								
Presumed intent to cause harm (IH)								**Presence of Implicit Extremist Content (IHC)**
Normalise hateful/violent narratives (IH)	Propagate hateful/violent conspiracies (IH)	Foster hate/hostility towards an out-group (IH)	Trigger to take online harmful actions - e.g. harassment/doxxing (IH)	Trigger to take offline harmful actions (IH)	Other (IH)	Other (specify)	**Presence of IH**	

TEST - Step 3 Threshold			
Option 1 (1 indicator)	***Option 2 (at least 2 indicators)***	***Option 3a (CM + 1 other indicator)***	***Option 3b (IH + 1 other indicator)***

Annex 7 : Assessment of the Platforms

Assessment of the Platforms									
	Instagram	Reddit	Snapchat	Tiktok	Facebook	Discord	Youtube	Telegram	Twitter/X
Extremist Content (Jih/RwX/ borderline)	Jih/Rwx/ borderline	RwX/ borderline	borderline	Jih/RwX/borderline	Jih/RwX/borderline	Jih/RwX	RWX/ borderline/ historically a lot of jihadi content	Jih/RwX	RWX/ borderline
Number of users/ followers globally per month	2 billion/ 259 mio in EU	1.2 billion/ 15.9 mio users in EU	750 million/ 102 mio in EU	1.6 billion/ 135.9 mio in EU	3.07 billion/ 259 mio in EU	196.2 million/ claims it has less than 45 mio in EU	2.50 billion/ 416.6 mio in EU	950 million/ claims that it has less than 45 mio in EU	4 billion/ 115.1 mio in EU
Average age of users	twenties to early thirties	mid-twenties	catering to young audience	catering to young audiance	mixed audiance	as young as 16. on average mid-twenties	mixed audiance	22-44 years	predominately (74,3%) male, largest age group is 25-34
Type of content (text/video/ memes/ photos/ other)	images and text-based comments; messenger function	text largely in chat format; memes	temporary images and text-based messages	video clips and text-based reactions	images/videos and text; messenger function	Gaming with textbased interactions	videos with text-based comment section	text-based chats through which viedos and images can be shared	text, videos, images
Relevance for the Netherlands	“high (8 million users)”	Medium	Medium	Medium	High	Medium	High	High	High

Who can make moderation decision (platform or users themselves)	platform	platform and users: moderators of subreddits are users who can moderate content in that specific subreddit. Depending on the size of the subreddit, that can be several people	platform	platform	platform	platform	platform	platform but minimal	platform but minimal these days
Amount of content already removed by platforms themselves					90% (through AI identification), see The Future of Free Speech, Preventing "Torrents of Hate" or Stifling Free Expression Online? An Assessment of Social Media Content Removal in France, Germany, and Sweden (Nashville: The Future of Free Speech, 2024), 53, https://futurefreespeech.org/wp-content/uploads/2024/05/Preventing-Torrents-of-Hate-or-Stifling-Free-Expression-Online-The-Future-of-Free-Speech.pdf.		99,5% (through AI identification), see The Future of Free Speech, Preventing "Torrents of Hate" or Stifling Free Expression Online? 53.		

Size/ location of platforms	Ireland (Belongs to Meta)	Netherlands	Netherlands	Ireland	Ireland (Belongs to Meta)	Netherlands	Ireland (Belongs to Meta)	Virgin Islands/ has legal representative in Belgium	Ireland
Means used to detect harmful content	AI	users can flag it; combination of automated and human moderators		Recently changed to mostly detecting by AI (80%), see Types Digital, "AI Takes Over Content Moderation at TikTok," Medium, published October 17, 2024, https://medium.com/@types24digital/ai-takes-over-content-moderation-at-tiktok-d4131bc6ae74.	users can flag it/ AI	AI	AI	platform	platform; users can flag it

Easy accessible for OSINT	Medium	Yes	Medium	Medium However, additional tools might be needed: OSINT research on TikTok is complicated by algorithm-driven content, and limited search functionalities, which are often deleted or difficult to trace. However, researchers have developed methods to extract user data and analyse profiles. For instance, utilizing TikTok's API and third-party tools enables the collection of user information and content for analysis. See "A Guide to OSINT Investigation and Research on TikTok," Steve Adams, Intelligence with Steve, published November 13, 2019, https://www.intelligencewithsteve.com/post/osint-tiktok.	Medium The ability to conduct OSINT on Facebook is limited by private groups and encrypted messaging features. Moreover, Facebooks algorithms determine what content is visible to users, and borderline content might be deprioritized or removed altogether before researchers can even get access to it. However, advanced keyword monitoring in public groups or resources like the Social Media OSINT repository provide tools for extracting and analysing data from Facebook. See ""The-Osint-Toolbox / Social-Media-OSINT,"" CQCore, and Roshan Behera, GitHub, accessed January 21, 2025, https://github.com/The-Osint-Toolbox/Social-Media-OSINT?tab=readme-ov-file#faceboo.	No	Yes	No Telegram's private and encrypted channels pose significant challenges for OSINT research. The Telegram OSINT Toolbox offers resources to search leaked data and analyse user information (GitHub), facilitating the monitoring of extremist activities. However, other issues often arise, such as extremist groups frequently changing channel names or invite links to avoid detection. Even when content is removed, it tends to resurface in duplicate or mirrored channels, making it harder to track and address effectively.	Medium Twittser/X is known to not be supportive of research anymore since Musk took over. For example, researchers are vetted and must pay a fee to access the API which was freely accessible before. Other OSINT features have also been deactivated.

Contactless research possible Y/N	Yes	Yes	Yes	Yes	Limited extent some interesting groups might be closed, contact would be required to access those	No	Yes	Yes	Yes

About the Authors

Dr Bibi van Ginkel, LLM is the Programme Lead of the Preventing and Countering Violent Extremism (PCVE) pillar of ICCT and Senior Research Fellow. She also manages ICCT's Centre of Excellence on Monitoring and Evaluation. Her expertise spans a wide variety of focus areas of ICCT, from rule of law issues to trends and threat developments, counterterrorism, countering violent extremism responses, and prevention strategies, and the nexus between development and PCVE.

Mrs Tanya Mehra, LLM is the Programme Lead of the Rule of Law pillar of ICCT and Senior Research Fellow. With a background in international law, her research focuses on rule of law approaches in countering-terrorism. Her main areas of interest are international (criminal) law with a focus on bringing alleged terrorists to justice for the full range of crimes they have committed in full compliance with human rights.

Ms Merlina Herbach, LLM is a Research Fellow for the Rule of Law pillar of ICCT. With a background in international law and political science, her work focuses on rule-of-law and human rights compliant responses to terrorism and violent extremism. At ICCT, Herbach is further responsible for the maintenance of two flagship databases, the Foreign Terrorist Fighters Knowledge Hub and the Interlinkages Database.

Mr Julian Lanchès, **MA** is a Junior Research Fellow for the Current and Emerging Threats Programme at ICCT. He specialises in the far right and Islamism, with his work focusing on different manifestations of extremism and terrorism online, the role of disinformation, propaganda, and identity in violent radicalisation, terrorist actors' strategies and modus operandi, as well as the mainstreaming of extremism.

Mr Yael Boerma, BSc is Research Assistant for the Preventing and Countering Violent Extremism pillar of ICCT. He holds a BSc in Political Science from Radboud University. Currently, he is completing his MSc in Political Science, specialising in International Relations and Political Theory.

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