Borderline Content

Understanding the Gray Zone

GIFCT
Global Internet Forum to Counter Terrorism

Dr Erin Saltman,
GIFCT, Director of Programming and Partnerships
Micalie Hunt,
GIFCT, Programming and Partnerships ssociate

Background

This GIFCT paper on borderline content is a contribution to a broader European Union Internet Forum (EUIF) handbook on the subject. The EU Internet Forum's key objective is to work hand-in-hand with the industry, EU Member States, Europol and civil society to provide instruments and concrete actions to jointly prevent the dissemination of harmful content leading to offline violence. The EUIF brought together research and perspectives from experts, tech platforms, and member states with the purpose of providing non-legally binding guidance on how to better understand and respond to borderline content that may lead to radicalization and violent extremism. Its objective is to understand the line between this content and violent extremism, but not to provide guidance on the borderline of the legality of such content under EU or national laws.

This handbook is the result of multi-stakeholder exchanges within the EUIF, in which all parties agreed on the need to provide support to tech companies on how to identify and limit the spread of borderline content that can lead to violent extremism and terrorism. dditionally, the EUIF and partners firmly believe that in the context of addressing borderline content in relation to radicalisation and violent extremism, any measures need to be based on fundamental and human rights. In view of online content that is linked to extremism and hate speech, any measures by platforms should be undertaken without unduly affecting the freedom of expression and of information of recipients of the service. ll the analysis and information contained in the wider handbook was provided by EU Institutions, GIFCT, EUIF Member States, civil society organisations, as well as key external stakeholders and researchers with a strong expertise in this field.[1] This report presents GIFCT's contribution to the EUIF research and analysis on this topic.

Introduction

The term "borderline content" has increasingly come up in discussions about processes of radicalisation leading to violence. This catch-all term has become prevalent in dialogues convened by the EU Internet Forum (EUIF), Global Internet Forum to Counter Terrorism (GIFCT), and the Christchurch Call to ction (CC). In these forums governments, technology companies, and expert stakeholders work to understand and develop action plans on the most pressing issues and efforts to counter terrorism and violent extremism online. 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 and violent extremist activities. Given the calls to build out processes that address borderline, it is important to provide a more nuanced understanding to the types of content that fall under scope of 'borderline content.' If the parameters of borderline content can be better defined, stakeholders will be able to better identify what potential actions can and should be taken to mitigate the risk of online harms at the periphery of terrorist and violent extremist exploitation online.

1 Specific contributors included - EU Services contractors: DG JUST, DG CNECT, the Fundamental Rights gency (FR), EU External ction Service (EE S), EU IRU at Europol, The Radicalisation wareness Network (R N); EU Member States: Germany, Italy, France, the Netherlands, Romania, Czech Republic, Spain, Media Council Slovakia; and Civil Society Organisations and think tanks: Tech gainst Terrorism, Christchurch Call dvisory Network (CC N), Institute for Strategic Dialogue (ISD), Center for Countering Digital Hate (CCDH), GNET, Counter Extremism Project (CEP), lexander Ritzmann, and Lisa Kaati.

Bringing borderline content to the fore of multi-stakeholder debates in and of itself highlights that this sector has advanced significantly. Previously, cross-sector forums convened to highlight the most obvious examples of terrorist exploitation online. However, as efforts by GIFCT member companies and tech companies willing to come to the table have evolved, so too has a more nuanced discussion about content that is harder to define, but seems within the realm of scrutiny for wider efforts to combat radicalising influences towards violence. This GIFCT contribution to the EU Internet Forum discussion on borderline content aims to give parameters to the term itself and provide better understanding of the relevant online policies and practices GIFCT member companies are taking in relation to what might be considered borderline content.

Variations in Tech pproaches to Borderline Content

For technology companies and their relevant platforms, national laws and government legislative guidelines exist to compel the removal of illegal content. These legislative frameworks also create the legal processes for the potential disclosure of data by tech companies to legally mandated government bodies where appropriate. bove and beyond illegal content, technology companies are often tasked with developing platform guidelines and policies for users that dictate what content and actions are acceptable on their platform. The capacity for a platform to develop nuanced policies or tooling to facilitate policy actions depends greatly on four things;

1. The human resources with subject matter expertise that a platform is able to hire,
2. The engineering and tooling support a platform is able to give to a harm type,
3. The awareness or prevalence of a certain harm type on the platform,
4. The external pressures by government, media and civil society pressuring a company to prioritise focus on a certain online harm issue.

Most global technology companies, depending on the tools and user experience a platform offers, have to think through online parameters for the culture they want to build for acceptable behaviour and consequences for users if they cross those lines. This is not dissimilar to how national and international governments think through legal frameworks for citizens. However, given the scale and global nature of online users and content, there will always be trade-offs between human and technical resources in relation to which policy areas demand prioritisation. There are high prevalence violating activities with relatively low real world harm risks (like non-scam related spam) and there are low prevalence violating activities with high risk for real world harm (like terrorist and violent extremist exploitation).

Even in cases where one company owns many different platforms, these platforms might have different policy lines around the topics that make up “borderline content”. This is because the surfaces and functionality a platform provides, stated purpose of the platform, and visibility of harms signal on each platform may differ. Having different degrees of variation for policy actions taken on borderline content is neither inherently good or bad. Some platforms, for example, might have a much lower

threshold for removing violent or graphic content because the platform is meant to cater to more professional networking. Whereas other platforms might protect wider speech and expression, knowing that their platform is used by activists, journalists, and marginalised communities who share a range of socio-politically and contextually sensitive content. Other platforms might have fewer content-focussed policies because the platform architecture prioritises user privacy, giving less content-visibility to moderators, such as in end-to-end-encrypted spaces.

GIFCT is not calling for a unified definition or criteria of actions against borderline content, but is calling for a better contextual understanding of the sub-categories of policy areas that make up the term and what actions might be available for tech companies. Given that borderline content needs 'borders,' and these borders differ across platforms, political contexts, and geographical contexts, the efficacy of any one company's approach to be utilised as a cross-platform example is limited. In turn, content that is permissible on one platform may be flagged and addressed as prohibited or limited borderline on another. Therefore, it is worth reviewing where more alignment in policy and practices can be pushed for when it comes to contentious borderline content, and where policy makers should continue to recognise protected speech and the values upheld in democratic countries.

What is Borderline Content in Relation to Terrorist & Violent Extremist Content?

Borderline content can be, and has been, conceived of in two ways. cademics and researchers tend to refer to borderline content as content usually protected by free speech parameters in a democratic environment, but inappropriate in public forums ie. "borderline illegal", or "lawful but awful" (Heldt, 2020). Tech companies tend to speak about borderline content as content that brushes up against a platform's policies for violating content ie. "borderline violative"(YouTube, 2019; Meta Transparency Center, 2023) but is not clearly violating a policy. Importantly, the literature on borderline content is overly reliant on tech platform definitions without a corresponding inquiry into how the term should be defined and what types of content would or should fall into scope (Murray, 2021; Bell, 2022; Gillespie, 2022).

These two categorisations for borderline content are related. It is broadly agreed that although borderline content is not technically illegal, it still has the potential to cause harm. Subsequently, there is pressure for tech companies to better understand and take appropriate action on this type of content, whether that is by removing it, taking other moderation actions, or ensuring it does not receive undue algorithmic optimisation reaching mass audiences. While democratic governments have deemed that certain segments of speech should be legally protected through the creation of legal frameworks, tech companies have recognised the harms that can arise from speech that is legal but problematic and harmful in the context of a particular public debate. Tech platforms, therefore, largely address any 'borderline illegal' TVE content through their specific terrorist and violent extremist or dangerous organisation policies.

However, the second type of borderline content, 'borderline violative,' also needs to be taken into consideration. Importantly, while 'borderline violative' TVEC content may not violate TVE policies,

such content may be actioned upon under other policies mitigating wider related harms. To better assess the state-of-play on how tech companies are addressing borderline TVEC, GIFCT outlined the primary online content and policy areas of its member companies that are most often associated with such content.[2] These include: hate speech; anti-refugee sentiment; stereotypes and dehumanisation; symbols/slogans and visual indicators associated with VE groups; meme subculture; misinformation; incitement to violence; anti-immigrant; weaponry/instructional material; violent, graphic, gory content; populist rhetoric - nationalism; anti-government/anti-EU; anti-elite; and political satire. While these categories cover content that may be described as 'borderline violative' TVEC, it is essential to note that each subcategory will have its own borderline violative content. This paper focuses on borderline content and sub-theme policy areas in relation to TVEC only.

These sub-themes have come up in ongoing conversations at the EU Internet Forum, within GIFCT thematic Working Groups,[3] and within the CC . These content policy areas have potential overlap or relations to terrorist and violent extremist content and wider processes of radicalisation. However, these topics exist far above and beyond the scope of terrorism and violent extremism in many ways that have no relation to TVEC or processes of radicalisation. s a reminder, research has shown time and time again that there is no one causal factor to an individual radicalising towards violence, and many have argued that the passive consumption of terrorist or violent extremist related materials plays a minor role in the overall process of radicalisation (Kenney, 2010; Reynolds and Hafez, 2017; Reiger et al, 2019, Lakomy, 2019). ny measures affecting content moderation policies and legislation should consider human rights objectives in ensuring that actions taken are legal, proportionate, and defendable. GIFCT emphasises that 'borderline content' should not be considered a moral designation and that such content may have legitimate uses in a number of circumstances beyond its utilisation by terrorists and violent extremists.

Building off of academic insights by the Global Network on Extremism and Technology (GNET), borderline content types can be mapped onto the following policy areas and TVE tactics (McGuffie, 2021) while also being utilised for non-TVE uses.

2 See GIFCT Official Website on Membership: https://gifct.org/membership/

3 See GIFCT Official Website on Working Groups: https://gifct.org/working-groups/

Borderline Content Sub-Theme Policies and Use in TVE and Non-TVE Cases

Borderline Content Type (Policy Category)	TVE Tactics + Content Types	Non-TVE Uses or Content
Violent Content, Graphic Content, Gore	"Promotion and glorification of violent extremism including references to and descriptions, photos, and videos of extremist violence and hate crime"	Journalism and reporting on situations, active conflicts, as well as academic research and educational sharing of content for analytical purposes
Weaponry + Instructional Material		
Symbols + Slogans and Visual Indicators ssociated with VE Groups	Using whistle symbols, emojis, and coded language to evade moderation efforts to remain on a platform and signal like minded users	cademic research and identification, news articles and reporting, as well as broader symbols, visual iconography, and numeric indicators being used in non TVE settings
Meme Subculture	Creating a sense of collective identity and internal group cohesion through 'secret' messaging only an in-group is aware of, evading moderation because it is a 'joke' or through the confusion of visual media	Culturally-relevant humour/messaging and communication as well as offensive joking outside of TVE contexts
Incitement to Violence	Promoting or inspiring attacks and intimidating online audiences in advance of intended offline action	Incitement to violence happens in a range of sociopolitical climates and scenarios that might violate a policy but are unrelated to TVE incidents or activities
Hate Speech	"Promotion of hate-based beliefs, ideologies, and discrimination. Facilitating scapegoating and false attribution of societal ills. empowering sympathisers with misguided sense of superiority"	Hate speech, bullying, harassment and threats happen in a range of sociopolitical climates and scenarios that might violate a policy but are unrelated to TVE incidents or activities
Bullying, Harassment, and Threats		
nti-Refugee / Immigrant Sentiment	Provides scapegoat for societal ills, solidifying an in-group and empowering sympathisers to feel superior to outgroups	Critical and even overtly antagonistic dialogues around refugee and migrant scenarios tend to be part of wider normative political discourse by political figures and are discussed openly on many mainstream media outlets
Stereotypes and Dehumanisation	Solidifies the antagonism towards a defined out-group targeted by TVE groups, allowing for consolidated demonisation of a perceived "enemy"	Stereotyping based on protected categories of people and dehumanising language happens in a range of sociopolitical climates and scenarios that might violate a policy but are unrelated to TVE incidents or activities
Mis- and Disinformation	"Provides false but appealingly simplified explanations within circumstances that create fear and societal uncertainty," often solidifying an outgroup or enemy of a TVE ideology	Mis and disinformation is spread widely by non TVE actors, far above and beyond processes of radicalisation, often spilling into mainstream discourse, espoused by political figures, and mainstream media.

Populist Rhetoric - Nationalism	Scapegoating and subjugating particular outgroups (identifying who is and is not of that national identity) empowering supporters to feel superior	Legal mainstream nationalism and populism, also instituted and espoused by both fringe and mainstream political entities
nti-Government/ nti-EU nti-Elite	Scapegoating and subjugating particular elite outgroups eroding trust in due process to promote alternative means for empowerment or change	Voice legitimate frustrations with socio political and economic situations under free speech protections
Political Satire	Create sense of collective identity and internal group cohesion, avoid censorship based on delivery as a 'joke'	Humorously draw attention to political situations - votes, elections, politician gaffes

This table quotes and builds off of the work originally presented in a GNET insight by McGuffies (2021).

Researchers and practitioners in this field continue to see the adversarial shift that when targeted policies increase on definable terrorist and violent extremist content (TVEC), bad actors decrease overt violating speech on that platform and replace it with 'borderline violative' content to evade moderation. Terrorists and violent extremists are aware of platform policies that may decrease their ability to disseminate particular forms of content. ccordingly, these actors often knowingly produce content that comes close to, but does not violate, existing platform policies ie. 'borderline violative content.' The Global Network on Extremism and Technology (GNET) has produced a number of Insights that demonstrate how borderline content can function as a strategy that some TVE entities employ to evade detection or restrictions in online spaces.

1. Borderline content allows for the 'softening' and mainstreaming of extremist beliefs to avoid censorship and moderation (Won and Lewis, 2021).
2. Populist racially and ethnically motivated extremist groups operationalise borderline content on social media to dilute their message and pursue increased recruitment of 'non-aligned' individuals (llchorn, 2021).

Experts acknowledge it is necessary to provide parameters to the term (Rogers, 2022). Given the adversarial nature of terrorism and violent extremism online, it is important to acknowledge that borderline content likely will not be accurately addressed by a static set of parameters. Content that is considered inappropriate or borderline changes in different political, cultural, and temporal contexts, which must be taken into account when attempting to take actions on this type of content.

ddressing Borderline Content - Possible ctions

Despite the absence of a common definition on borderline content, the outlined borderline sub themes listed above make-up online policy areas that many GIFCT members have proactively taken steps to address. The Trust and Safety Professional ssociation (TSP) has outlined the various moderation actions companies are currently taking across a range of borderline content

policies: content deletion, banning, temporary suspension; feature blocking; reducing visibility either by removing from recommendations, downranking, or auto-collapsing comments; labelling; demonetization; withholding payments; and referral to law-enforcement (See TSP Reference). Referral to law-enforcement is unlikely to apply to borderline content, as any content that necessitates law-enforcement referral would be violative of a tech platform's policies and as such would not be classified as 'borderline.'

Enforcement ctions of Tech Companies in Moderation Efforts

Enforcement ction	Definition
Content Deletion	Removing content that violates a platform's policy. Most common action taken by platforms.
Banning	"Permanent removal or blocking of a user or account from the platform. May include banning any new accounts that the user attempts to create or uses to access the platform. Content from before an account is banned may still be visible or may be removed as part of the ban."
Temporary Suspension	"Identical to banning, but lasts only for either a specified period of time or until the user completes certain specified actions, after which account is automatically reinstated." May be used as a precursor to permanent ban.
Feature Blocking	"Encompasses any restriction of access to certain features of a platform based on previous actions of a user, either temporarily or permanently. Might involve removing access to features that had been misused in the past, or to features that would be considered higher risk or more difficult to moderate, such as live streaming. llows users to remain active on the platform, while minimising potential harm from their actions".
Reducing Visibility	"Refers to steps that reduce how often and how prominently a piece of content or an account is viewed. These steps are most often used on platforms in which the product itself guides and curates a user's experience with algorithms. This may include removing the user/content from features such as recommendations or trending stories; downranking the user's or content's position in search results or feeds; and auto-collapsing comments on threaded posts."
Labelling	"Involves attaching a message to a user or piece of content to provide information to the viewer. These labels can be used to inform the viewer of any concerns or of important information relevant to the content or topic discussed."
Demonetization	"Prevents users from earning income and specifically applies to platforms where users can earn money from their content, usually through advertising. Demonetization is often applied to content that is allowed on the platform, but which is controversial or which advertisers may not wish to sponsor or be directly associated with."

The contents of this table and quotes are taken from the TSP mapping for Enforcement Methods and ctions (see bibliography).

Importantly, it should be understood that while all GIFCT member companies, and tech platforms more broadly, can draw from the same list of enforcement methods and actions, the specific actions utilised are constrained by the characteristics of a particular tech platform and resources. How robust and nuanced the response to these policy areas are depends on the previously listed four-point criteria and availability of; human resourcing, engineering and tooling support, awareness of harm types, and external pressures from other sectors. There are both proactive and reactive human and tooling resources that can be employed to review and take action on the content making up borderline content sub themes. The more the content related to a policy topic can be clearly defined and is clearly harmful, particularly relating to real world harms, the more likely proactive tools and resources can be easily deployed without the fear of over censorship.

Smaller and less resourced platforms will be limited in their ability to undertake enforcement actions that require more nuanced moderation or technical expertise. Further, these smaller platforms may be unable to proactively contribute to the policy debates on what types of content should be moderated and may be more likely to follow the lead of their larger, more established counterparts. However, the following efforts are all aimed at facilitating support for smaller companies being able to advance their efforts on clear TVE content and some of what is still considered borderline in terms of illegal content as determined by regulators:

Tools to Support Smaller Tech Companies in Countering TVEC

Effort	Description
GIFCT's Hash Sharing Database	GIFCT's Hash-Sharing Database enables GIFCT member companies to quickly identify, and share signals, of terrorist and violent extremist activity in a secure, efficient and privacy-protecting manner. Known as perceptual hashes, a hash is a numerical representation of original content (video, image, PDF, or URL) that cannot be easily reverse-engineered to recreate the content. These hashes are added to the database with a series of labels corresponding to the GIFCT database taxonomy[4] to help other members understand what content corresponds to the hash, including content type, terrorist entity that produced the content, and its behavioural elements. GIFCT member can then select a hash to see if it identifies and matches to visually similar content on their platform.
GIFCT & Faculty.ai Terrorist Classifiers	Faculty has begun work with GIFCT to widen access to terrorism moderation tooling for smaller content hosting platforms to provide a delivery model that facilitates small platforms' access to terrorism classification models at no cost. This will offer small online platforms who are members of GIFCT free access to a suite of advanced I models that developed over the past five years to classify Daesh and al-Qaeda propaganda in multiple formats with exceptionally high performance (Drew 2023).
Meta's Hash Matcher ctioner	Meta has made available a free open source software tool it has developed that can help platforms identify copies of images or videos and take action against them en masse (Clegg, 2022). Hasher-Matcher- ctioner (HM) can be adopted by a range of companies to help them stop the spread of terrorist content on their platforms, particularly of potential use for smaller companies lacking resources (Facebook/ThreatExchange, 2022). HM builds on Meta's previous open source image and video matching software, and can be used for any type of violating content, including to counter terrorist and violent extremist content.
Jigsaw and Tech gainst Terrorism's moderation tool	Being developed by Google's research and development unit, Jigsaw, in partnership with Tech gainst Terrorism and with support from GIFCT, this tool aims to help human moderators make decisions on content flagged as dangerous and illegal. Testing and development will continue throughout 2023 (Criddle, 2023).
Tech gainst Terrorism's TC P	The Terrorist Content nalytics Platform (TC P) seeks to disrupt terrorist use of the internet by facilitating the quick and accurate removal of terrorist content (See TC P Official Site). It does this by alerting terrorist content to tech companies when found on their platforms. team tracking terrorist migration across a variety of tech platforms flag URLs containing terrorist content to the TC P. TC P sends alerts to tech platforms about terrorist content found on their sites and checks the status of URLs to determine when content is removed.

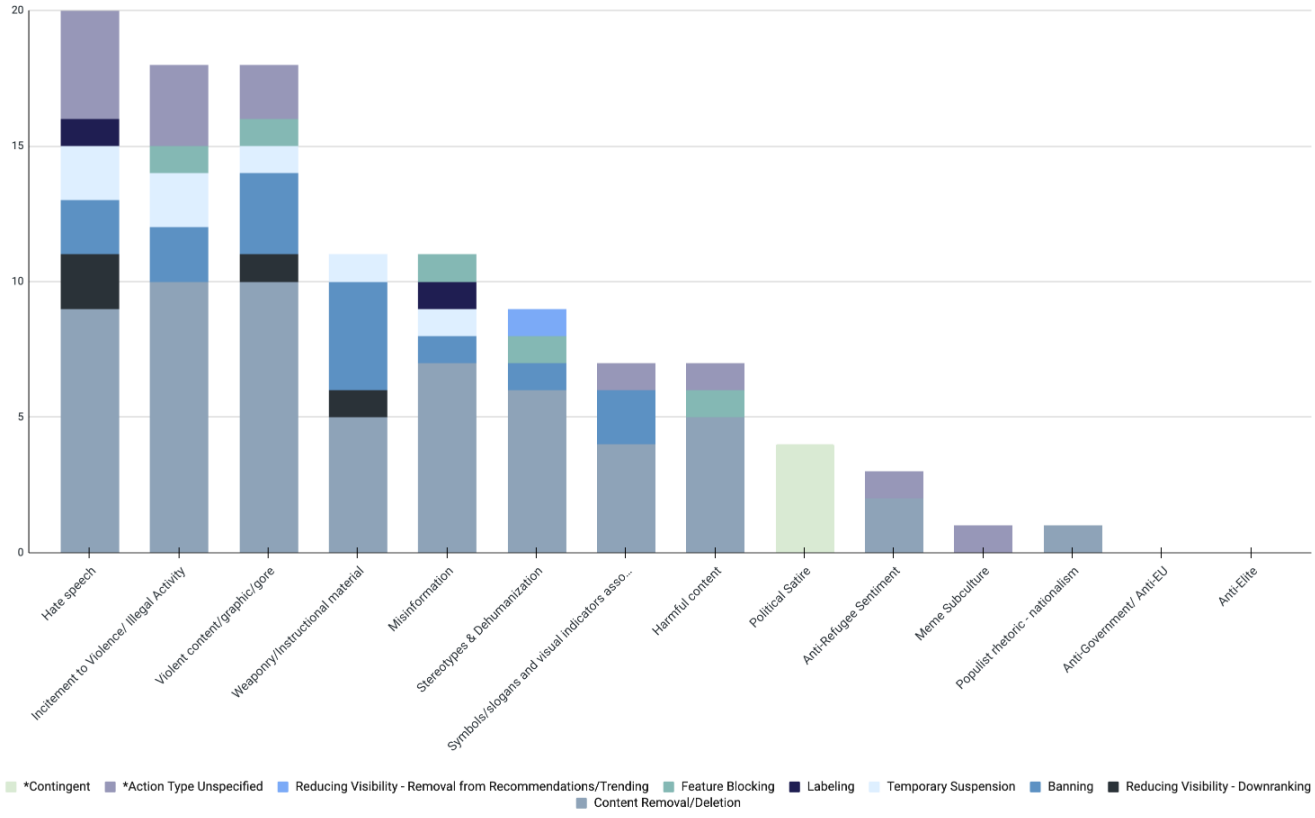
4 See the 2022 GIFCT nnual Transparency Report for detailed description of the current Hash Sharing Database taxonomy, pp. 22 - 36, https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

Comparatively, larger or higher risk companies may be better able to build more nuanced internal processes as well as having better capacities to create scaled partnerships for fact checking or trusted flagging. The type of enforcement method and action available to a tech platform is also dependent on the platform type. Platforms that do not rely on algorithms to curate a user's experience through surfaces like a "newsfeed" or "search results", are unlikely to utilise actions that reduce visibility either through downranking or removing content from recommendations. Further, platforms that do not contain user-generated content would be less content-focussed in removal or deletion policies. ccordingly, while some tech platforms are able to utilise a wide range of enforcement methods for borderline content others are constrained because of the nature of the enforcement options available. Thus, when considering how tech platforms are identifying and actioning borderline content, it is important to acknowledge that there will not be a one-size-fits-all approach or ability.

Reviewing Company Policies and ctions

GIFCT reviewed its 22 GIFCT member companies' policies against the 14 sub-themes identified as making up borderline content in relation to TVEC. ctions taken were drawn from the TSP enforcement methods and actions outlined above; however, two additional categories were added to provide greater nuance to our comparison. Policies that listed multiple potential methods or actions were categorised as 'action type unspecified.' Further, a 'contingent' action was added for borderline content subcategories that required multiple signals within a particular content type.

Borderline content types and enforcement action taken by GIFCT member platforms



The table shows where and how GIFCT member companies currently take action against TVE related borderline content sub-themes. These are taken from publicly available Terms of Service, User Guidelines, and public statements made by companies.

In analysing where current tech companies have policies and take actions, there are some areas of wider agreement and some areas with large variation. Variations occur both in where policies exist and what actions are applicable in those policy areas. The analysis provided six primary conclusions from the data.

1. **The more content ties to real-world harm, the more likely removals and remedial actions are prevalent.**

Broadly speaking, analysis showed that the more a borderline content policy area had direct links or associations with real world harm or off-line violence, the more likely it was that policies existed on a platform to take some form of remedial action against the content or user behaviour. This includes areas of incitement to violence, violent content, graphic content and gore. Relatedly but to a lesser extent, companies converged on policies for remediation on sale of weaponry, instructional material, and misinformation tied to real world violence.

2. **Policies where offline legislation and legal guidance is available correlates with where online policies have developed.**

The United Nations has a Strategy and Plan of ction on Hate Speech (United Nations, 2019), while the European Union has undertaken a number of efforts to act against hate speech (European Commission, 2021) both online and offline. Incitement to violence is also intimately linked to hate speech frameworks, as the combination of these two elements may make such speech illegal under rticle 20, paragraph 2 of the International Covenant on Civil and Political Rights (ICCPR) (OHCHR, 1966). ccordingly, when government and intergovernmental bodies create strong legal or academic frameworks for addressing specific types of speech offline content, tech platforms are able to follow their lead and create policies that action these specific types of borderline TVE content online. By following the lead of international bodies that have access to greater resources, tech companies can act on borderline TVEC while ensuring that human rights considerations are emphasised. dditionally a number of subcategories such as anti-refugee sentiment, harmful content, stereotypes and dehumanisation were subsumed under broader hate speech or graphic content policies for some platforms.

3. **Misinformation is difficult to define and action.**

Half of GIFCT's companies had established policies on misinformation. Some of these policies stipulated that they would only activate on election-related misinformation, while others sought to address misinformation more broadly. Given that categorising information as misinformation is often contextually-dependent and there is no agreed upon method for defining or definition of misinformation, some tech platforms may be hesitant to proactively create policies or may choose to deal with it through alternative measures.

4. **Content removal is the most likely tool for remedial actions.**

Tech platforms were more likely to remove or delete content and ban users than other types of enforcement actions. These two enforcement types are broadly applicable across a variety of tech platform types and are accessible to large and small companies alike, which renders them a first choice for tech platforms seeking to address borderline TVEC. Temporary suspension of user accounts, as the precursor to banning, is also broadly utilised. Bans and content removals also have a cross-platform impact, when content is removed from one platform it is unable to be shared or linked across to other platforms. However other types of mitigation, such as visibility reductions, may isolate the impact to a single tech platform as individuals can still share content to other platforms where they are not downranked.

5. **Nuanced enforcement is primarily only viable for large, well-resourced platforms.**

More complex and nuanced enforcement actions, including feature blocking and variations of reducing visibility, were less utilised by companies. These were mainly employed by larger, greater resourced companies. To note, that nuanced enforcement tools need both the tooling as well as the human resource to review incoming content picked up by tooling.

6. **Some borderline content areas are clearly protected by democratic principles on speech.**

ll GIFCT members protected user rights to be critical of governments or elites, in line with understandings of protected free speech in democratic environments. When government criticism or political satire was actioned, content needed to violate an additional existing policy to be actioned. Populist rhetoric, as it relates to nationalism, required the same additional criteria. Types of content that are heavily context dependent or ill-defined, including meme subculture, symbols, slogans and visual indicators associated with violent extremist or non-violent extremist groups were also less likely to be addressed by a policy and actioned. Memes in particular are incredibly context-specific and rapidly evolve, which makes these visual indicators hard to regulate. Since these categories are more difficult to link to offline violence or TVE and lack norms or policy frameworks, tech companies do not have a clear framework to ban these types of borderline TVE.

Conclusion

s dialogues about borderline content continue between governments, tech companies, and experts, it is important to understand the framing of the term, the policies and practices being advanced by technology companies, how government guidance and regulation plays a role, and where multi-stakeholder partnerships remain crucial.

Framing: The term "borderline content" is both subjective and manifold. It denotes a range of online policy or content areas that may have overlap with terrorist and violent extremist content or conduct, but are largely legal speech within democratic frameworks. Knowing that the term is used as an umbrella for a variety of sub-theme policy areas, it is important to understand that binary broad stroke statements demanding a particular action for all "borderline content" is not possible. s such, understanding the range of sub-themes and related online policies around those sub-themes is necessary.

Policies and Practices of Tech Companies: Looking at the sub-themes that make up TVE borderline content across GIFCT member company policies, it is clear that lots is already taking place in terms of moderation and remedial actions as outlined in this paper. nalysis of company policies found that the more a sub-theme policy area concerned content related to, or inciting real world harm, the more likely clear remedial actions were taken by tech companies. Overarchingly, the more broadly a sub-theme aligned with less defined "controversial" opinions or "lawful but awful" speech, the more speech was protected. In many cases tech companies are already going above and beyond clear legal guidance in taking actions on content. Looking at the range of tools available to take action on content, larger companies will continue to have more human and tooling resources to take nuanced approaches to borderline content.

Government Guidance: The more governments can define the TVEC related harm areas they are most concerned about, and the more this can tie to legal frameworks, the easier it is to encourage actions by tech companies in a principled manner. Even in cases where content is not removed, but is downranked or demonetised, there need to be principled policies behind the actions that are definable, defendable, and scalable. Governments should look to reflect on the sub-themes related to borderline content to better prioritise and scrutinise policy areas that are most directly tied to real world harm and existing offline policies.

Partnerships and Multi-stakeholder Efforts: GIFCT was founded with a multi-stakeholder approach to its governance and its work. Having diverse stakeholders working together is not just nice to have. It is paramount for success. Partnerships and multi-stakeholder efforts will continue to be crucial in (1) ensuring companies with less human or tooling capacities understand what adversarial shifts look like, and (2) are given the networks and tooling needed to develop cross-platform solutions. Countering terrorism and violent extremism online, including understanding the borderline content that might contribute to processes of radicalisation, relies on cross-sector collaboration to be effective.

BIBLIOGR PHY

llchorn, W., *Turning Back to Biologised Racism: Content nalysis of Patriotic lternative UK's Online Discourse*, GNET Insights, 22 February 2021, https://gnet-research.org/2021/02/22/turning-back-to-biologised-racism-a-content-analysis-of-patriotic-alternative-uks-online-discourse/

Bell, K., *YouTube could 'break' sharing on borderline content to fight misinformation*, Engadget, 17 February 2022, https://www.engadget.com/youtube-could-break-sharing-on-borderline-content-to-fight-misinformation-201819354.html

Clegg, Nick, *Meta Launches New Content Moderation Tool as it Takes Chair of Counter Terrorism NGO*, Meta Newsroom, 13 December 2022, https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/

Criddle, C., *Google develops free terrorism-moderation tool for smaller websites*, rs Technica, 3 January 2023, https://arstechnica.com/tech-policy/2023/01/google-develops-free-terrorism-moderation-tool-for-smaller-websites/

Drew, T., *How terrorists are capitalising on the cost of I*, Tech UK, 16 January 2023, https://www.techuk.org/resource/natsec2023-faculty-16jan23.html

European Commission, *Stop Hate: The Legal and Policy Framework in the EU*, European Commission Official Website, 2021, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/combating-hate-speech-and-hate-crime_en

Facebook/ThreatExchange, *Hasher-matcher-actioner*, GitHub, https://github.com/facebook/ThreatExchange/tree/main/hasher-matcher-actioner

GIFCT, *2022 GIFCT Transparency Report*, December 2022, https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

GIFCT, *Hash Sharing Database*, GIFCT Official Website, https://gifct/org/hsdb/

GIFCT, *Membership*, GIFCT Official Website, https://gifct/org/membership/

GIFCT, *Working Groups*, GIFCT Official Website, https://gifct/org/working-groups/

Gillespie, T., *Reduction / Borderline content / Shadowbanning*, Yale-Wikimedia Initiative on Intermediaries & Information, 20 July 2022, pp. 1 - 14 https://law.yale.edu/sites/default/files/area/center/isp/documents/reduction_ispessayseries_jul2022.pdf

Heldt, ., *Borderling speech: caught in a free speech limbo?*, Internet Policy Review: Journal of Internet Regulation, 15 October 2020, https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510

Kenney, M., *Beyond the Internet: Mētis, Techne, and the Limitations of Online rtifacts for Islamist Terrorists*, Terrorism and Political Violence 22, no.2, 2010, pp. 177–97.

Lakomy, M., *Let's play a video game: Jihadi propaganda in the world of electronic entertainment*, Studies in Conflict & Terrorism 42, no.4, 2019, pp. 383–406

McGuffie, K., *pplying Systematic Content Moderation for Extremist Deterrence*, GNET Insights, 2 November 2021, https://gnet-research.org/2021/11/02/applying-systematic-content-moderation-for-extremist-deterrence/

Meta, *Content Borderline to the Community Standards*, Meta Transparency Center, March 2023, https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards/

Murray, J., *What is "borderline" content on YouTube?*, Engage Web, 24 September 2021, https://www.engageweb.co.uk/what-is-borderline-content-on-youtube-19796.html

Rieger, D., L. Frischlich, and G. Bente, *Dealing with the Dark Side: The Effects of Right-Wing Extremist and Islamist Extremist Propaganda from a Social Identity Perspective*, Media, War & Conflict 13, no.3, 2019, pp. 280–99.

Reynolds, S.C. and M.M. Hafez, *Social Network nalysis of German Foreign Fighters in Syria and Iraq*, Terrorism and Political Violence 31, no.4, 2017, pp. 661–86.

Rogers, E., *The Role of User gency in the lgorithmic mplification of Terrorist and Violent Extremist Content*, GNET Insights, 21 September 2022, https://gnet-research.org/2022/09/21/the-role-of-user-agency-in-the-algorithmic-amplification-of-terrorist-and-violent-extremist-content/

Terrorism Content nalytics Platform, *bout*, https://www.terrorismanalytics.org/about

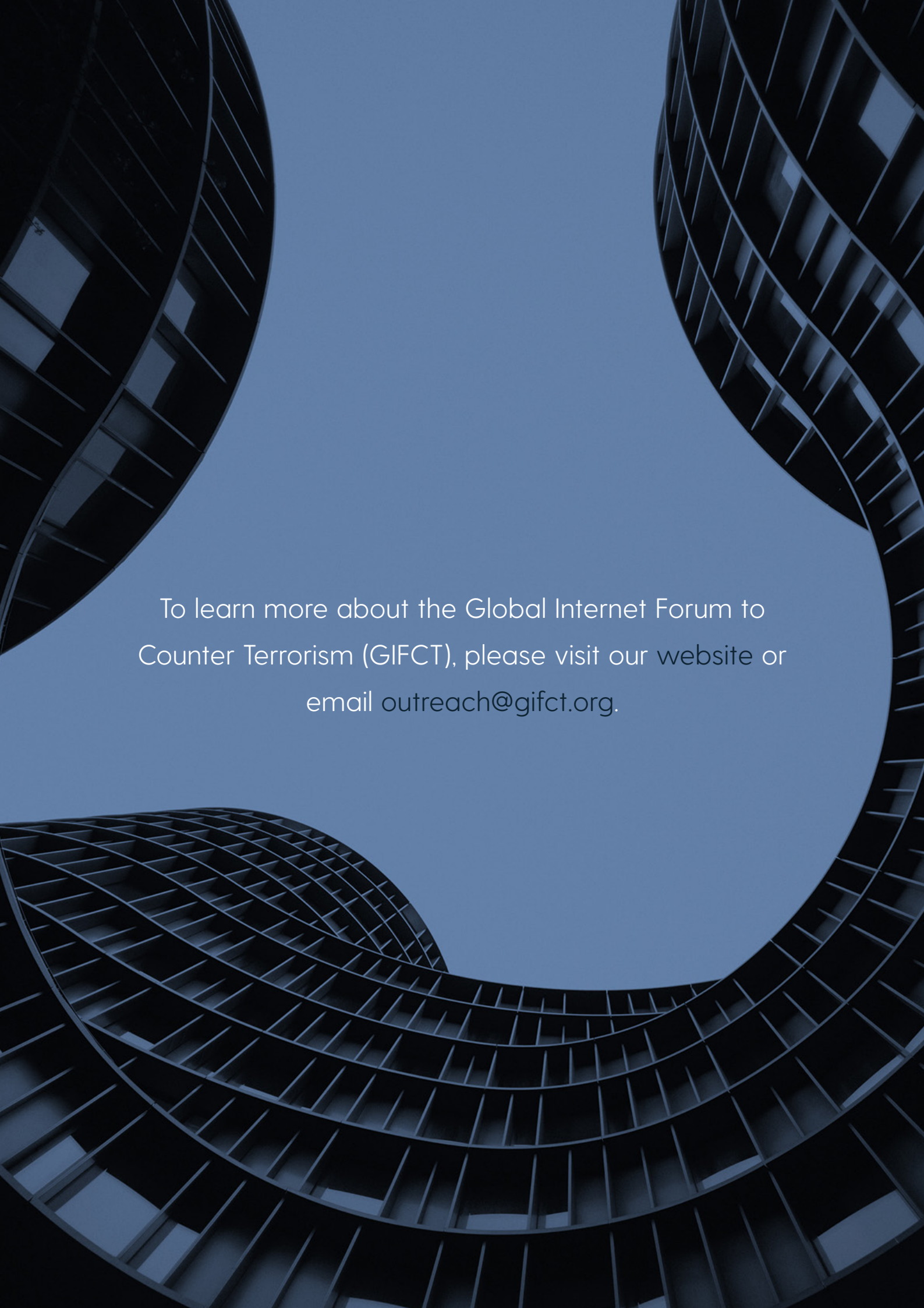
TSP , *Enforcement Methods and ctions*, Trust & Safety Professionals ssociation, https://www.tspa.org/curriculum/ts-fundamentals/policy/enforcement-methods/

United Nations, *United Nations Strategy and Plan of ction on Hate Speech*, May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/ ction_plan_on_hate_speech_EN.pdf

United Nations Human Rights Office of the High Commissioner, *International Covenant on Civil and Political Rights*, United Nations OHCSR Official Site, 16 December 1966, https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

Won, Y.B. and J. Lewis, *Male Supremacism, Borderline Content, and Gaps in Existing Moderation Efforts*, GNET Insights, 6 pril 2021, https://gnet-research.org/2021/04/06/male-supremacism-borderline-content-and-gaps-in-existing-moderation-efforts/

YouTube, *The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation*, YouTube Official Blog, 3 December 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/



To learn more about the Global Internet Forum to Counter Terrorism (GIFCT), please visit our website or email outreach@gifct.org.