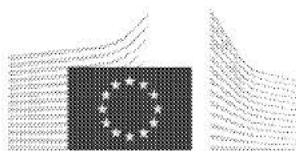


Section III:
EU Internet Forum Exhibits

Exhibit 28



EUROPEAN COMMISSION

DRAFT AGENDA
EU INTERNET FORUM
MINISTERIAL MEETING
hosted by
COMMISSIONER [Redacted]

8 DECEMBER 2021

HYBRID MEETING: IN PERSON (BERLAYMONT, SCHUMAN ROOM), AND ONLINE

INTERPRETATION:
EN-FR-DE-ES-IT-SI

14:30-14:40 *Family photo (for physical attendees)*

14:40-15:00 **Opening remarks (public session)**

- **[Redacted]** *European Commissioner for Home Affairs*
- **[Redacted]** *Minister of Interior of Slovenia, EU Council Presidency*

15:00-16:00 **“Ensuring security online, shared responsibilities”**

- **[Redacted]** *Europol on ensuring security online: changing landscapes and new trends of criminal activity online*
- **[Redacted]** *IN Internet Governance Forum, on the shared responsibilities for a safer internet: safety by design (industry) fundamental rights (policy), empowerment (civil society)*
- **[Redacted]** *Vice President Government Affairs and Public Policy for Europe, Google, on industries’ responsibilities to foster a more secure online environment*
- *Exchange of views*

16:00-17:00 **Topical debate 1: Fighting online child sexual abuse**

- *Commission response on preventing and combatting child sexual abuse*
- **[Redacted]** *Vice President European Government Affairs, Microsoft on safety by design*
- **[Redacted]** *Secretary of State for Child Protection, France on innovative approaches to safeguarding children online*
- **[Redacted]** *Chair, WeProtect Global Alliance, on new trends of child sexual abuse online*

- **Redacted** *e-safety Commissioner, Australia on safety by design (tbc)*
- *Interventions by Ministers, industry and other participants*

17:00 – 17:50 Topical debate 2: Countering Violent extremism and terrorism online – emerging challenges

- **Redacted** *Peace Research Institute Frankfurt on the use of video-gaming for recruitment and radicalisation*
- **Redacted** *Executive Director, Global Internet Forum on Counter Terrorism to address new trends on all types of extremism and challenges of industry in moderation*
- *Meta on initiatives to address borderline content and its impact on young users (tbc)*
- *Interventions by Ministers, industry and other participants*

17:50-18:00 Closing remarks

- **Redacted** *European Commissioner for Home Affairs*

Exhibit 29

From: [Redacted]@tiktok.com>
Sent: jeudi 24 mars 2022 20:21
To: [Redacted] (HOME); [Redacted] (HOME)
Cc: [Redacted] (HOME); [Redacted] (HOME); [Redacted] (HOME)
Subject: Re: RE: [External] Russia-Ukraine situation

Here are our recent initiatives on anti-semitism / authoritative content:
<https://newsroom.tiktok.com/en-eu/commemorating-holocaust-remembrance-day-eu>

[Redacted]

From: [Redacted]@ec.europa.eu>

Date: Thu, Mar 24, 2022, 20:12

Subject: RE: [External] Russia-Ukraine situation

To: [Redacted]@tiktok.com>

Cc: [Redacted]@ec.europa.eu>, [Redacted]

[Redacted]@ec.europa.eu>, [Redacted]

[Redacted]@ec.europa.eu>, [Redacted]

[Redacted]@ec.europa.eu>

Many thanks [Redacted] for this quick reply! I saw that this research got some traction in the news too, so would be good to know what your moderation team has found to be the issue.

It would be great to hear more about what you do on anti-semitism, as the anti-semitism action plan asks for more to be done on the online sphere.

[Redacted] is following that from our side.

[Redacted]

From: [Redacted]@tiktok.com>

Sent: Wednesday, March 23, 2022 10:42 AM

To: [Redacted]@ec.europa.eu> [Redacted]
[Redacted]@tiktok.com>

Cc: [Redacted]@ec.europa.eu> [Redacted]
[Redacted]@ec.europa.eu>; [Redacted]
[Redacted]@ec.europa.eu>; [Redacted]
[Redacted]@ec.europa.eu>

Subject: Re: [External] Russia-Ukraine situation

Hi [Redacted]

Apologies for the delay, I was travelling yesterday.

As regards the Newsguard report, I have a meeting on it with DG Connect later today. But in short, the researchers have not been in contact and we have been unable to emulate what they saw, so it is challenging to properly assess this. We don't believe that the experiment mimics standard viewing behaviour but in any case, we continue to respond to the war in Ukraine with increased safety and security resources as we work to remove harmful misinformation. We are partnering with independent fact-checking organisations to support those efforts and have submitted a couple of written reports to DG Connect on what we are doing, following their request to all platforms.

Regarding counter-narratives, this can be tricky on our platform due to the way that the TikTok For You Feed (FYF) works, which is very individualised and based on personal engagement. Our most advanced work in this area has probably been against antisemitism (we have worked with a number of partners) as opposed to violent extremism but in general, we are still trying to see how our platform can enable counter-narrative content. We are early in our journey overall and it is a challenging area in any case. We do some work against filter bubbles and to promote diversity of content in the FYF which helps but we are still examining how we would 'raise' certain voices.

Hope that helps, let me know if you have any further questions.

Kind regards

[Redacted]

From: [Redacted]@ec.europa.eu>

Date: Tue, Mar 22, 2022, 2:18 PM

Subject: [External] Russia-Ukraine situation

To: '[Redacted]@tiktok.com>

Cc: '[Redacted]@ec.europa.eu>, [Redacted]
[Redacted]@ec.europa.eu>, [Redacted]
[Redacted]@ec.europa.eu>, [Redacted]
[Redacted]@ec.europa.eu>

Dear [Redacted]

It was great to catch up the other day. I would like to come back with a couple of questions on what we see on Tik Tok related to Ukraine.

We saw research by Newsguard that suggested on Tik Tok users can stumble upon misleading content about the war in Ukraine within 40 minutes of signing up: Misinformation Monitor: March 2022 - NewsGuard (newsguardtech.com) I suppose you are aware and there are actions to avoid this from happening?

Another worrying development was the use of Russian TikTok influencers paid to spread Kremlin propaganda.

You may know that the EU Internet Forum also supports alternative narratives to those of extremists, and it would be good to know what your approach is to promoting alternative narratives, through civil society in particular. When we 'visited' the Transparency centre, You talked about Tik Tok's research on digital wellbeing and actions such as recommending different types of videos at night time. But do you have an approach to recommending different types of content to counter terrorist and violent extremist narratives?

Many thanks for any information and if it is quicker we can always have a quick call if you prefer.

Best wishes, [Redacted]

Exhibit 30

DIGITAL SERVICES

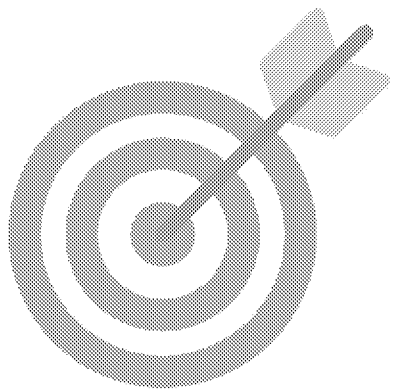
ACT & Algorithmic amplification

Redacted

EU Internet Forum

29 September 2022

OBJECTIVES



1. to create a **safer digital space** in which the **fundamental rights** of all users of digital services are protected

Spread of illegal content, sale of non-complaint goods and services

Protecting fundamental rights online

Tackling societal concerns such as disinformation and child safety

One set of rules across the entire EU single market

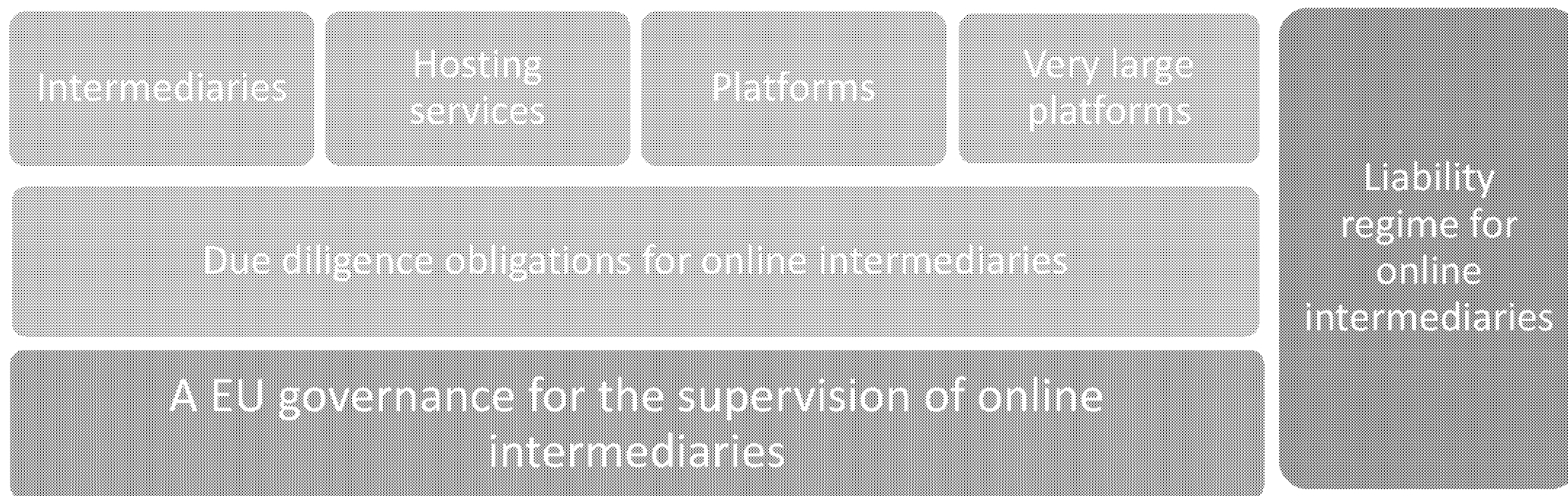
Proportionate, asymmetric obligations

Exemptions for Small and micro-enterprises

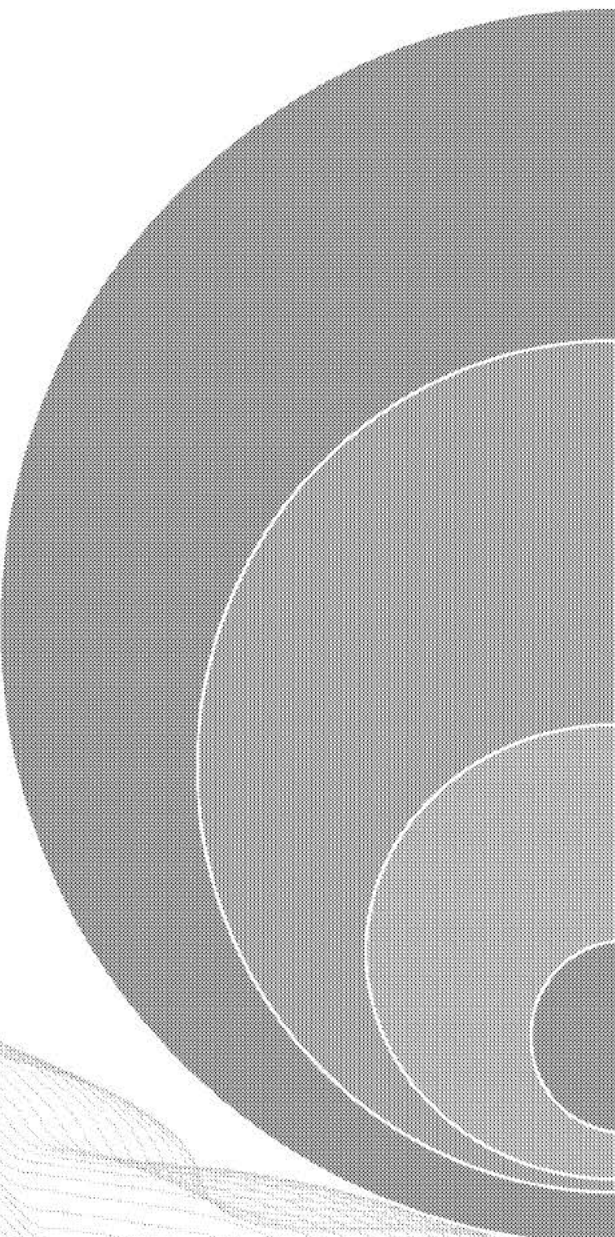
2. to establish a **level playing field** to foster **innovation, growth, and competitiveness**, both in the European Single Market and globally



The Digital Services Act

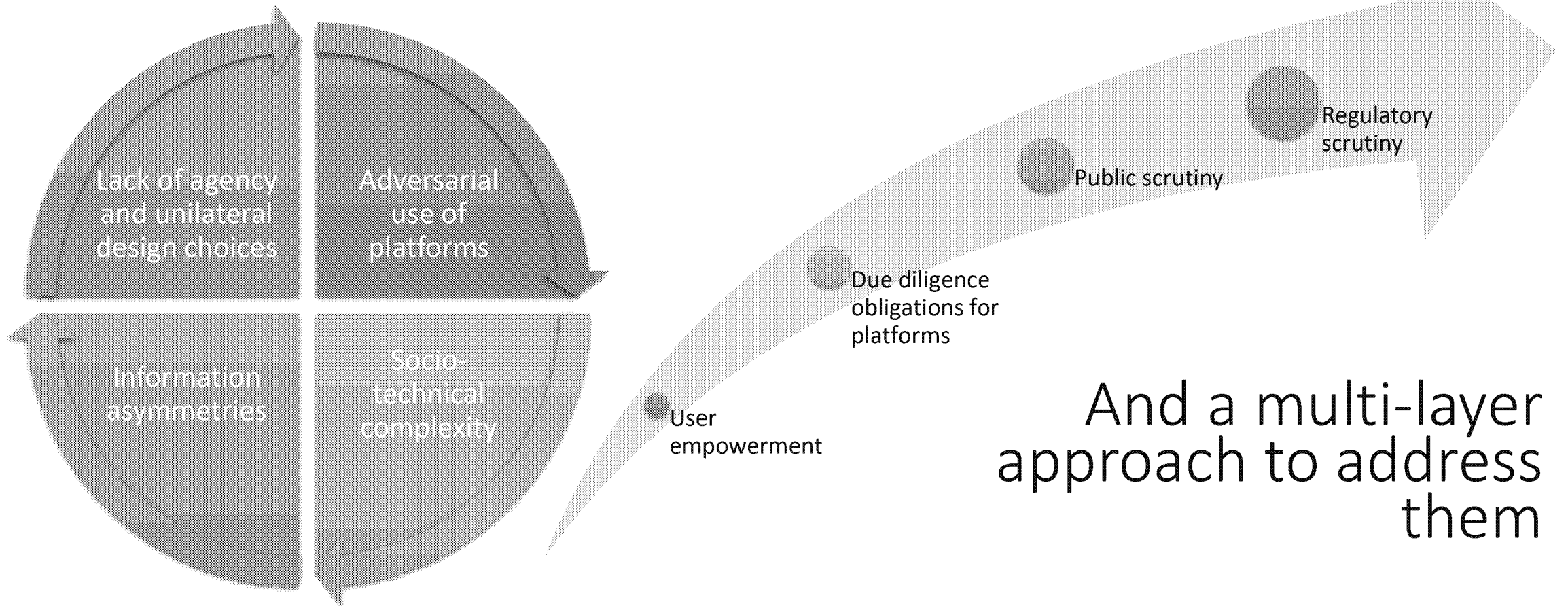


Due diligence obligations



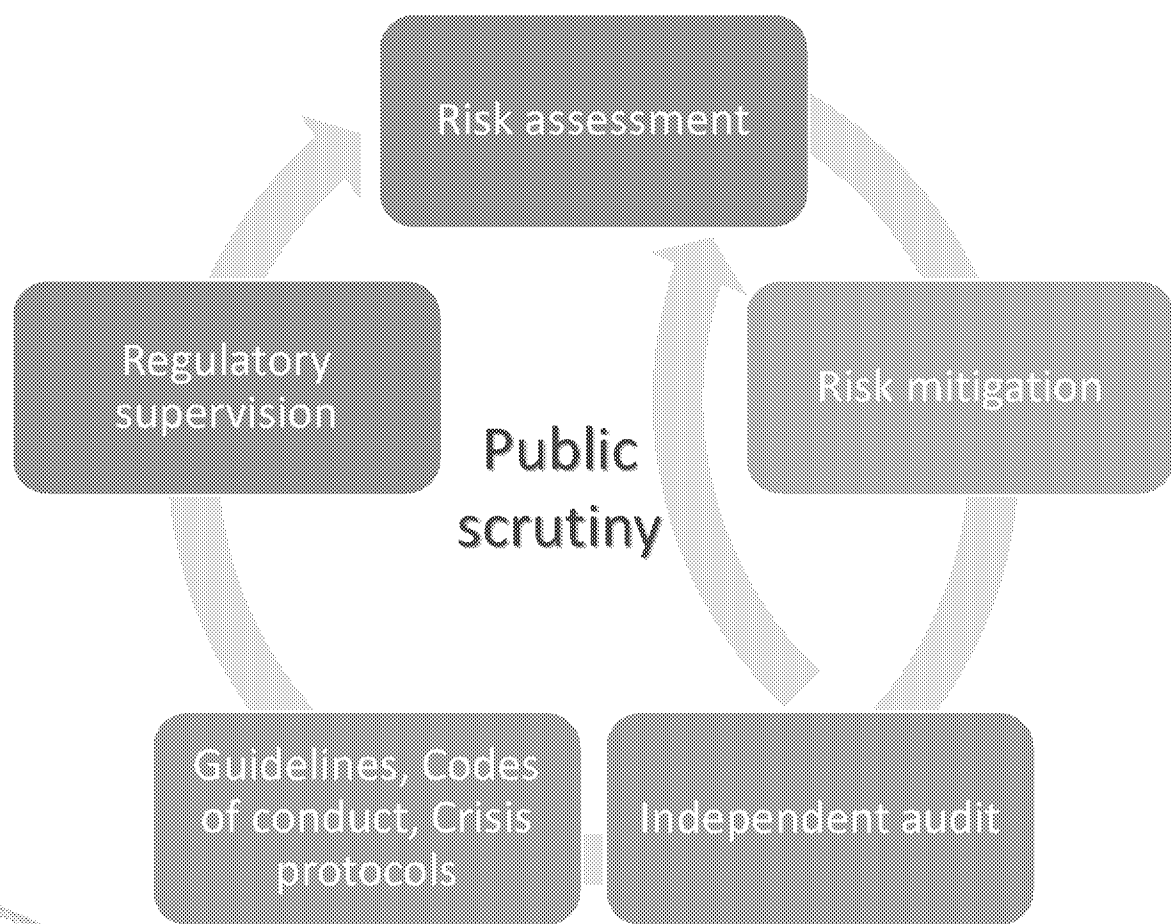
Very large online platforms	• Risk management, crisis response & audits • Recommender systems: choices • Ad repositories • Data access for researchers and supervisory authorities • Compliance officer • Further transparency reporting
Online platforms	• Internal & out of court complaint systems • Trusted flaggers • Limiting misuse • Obligations for marketplaces • Advertising transparency and bans on certain targeted ads • Transparency of recommender systems • Child protection measures • Bans on 'dark patterns' • Enhanced transparency reporting
Hosting services	• Notice & action • Information to notice-providers • Information to content provider • Suspicious criminal evidence
All intermediaries	• Points of contact & legal representatives • Clear terms and conditions & diligent, objective, proportionate enforcement • Transparency reporting

Core regulatory challenges in algorithmic amplification



And a multi-layer approach to address them

Supervised risk management



- **Adaptive regulation**
- A **dynamic** approach to identify and address **societal risks as they emerge**
- Covers the **use** but also the **core design of a service**, from its terms and conditions, to its algorithmic systems and optimisation choices

Assessing systemic risks

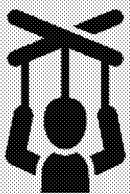
Obligation of means: reasonable, proportionate and effective measures



Dissemination of illegal content



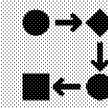
Negative effects for fundamental rights, including freedom of expression, data protection, privacy



Negative effects on other societal concerns: public health, security, civic discourse, electoral processes, mental and physical well-being, children

Iterative guidance and best practices reports

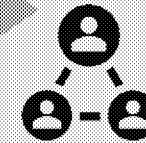
Measures may include:



Adapting content moderation or recommender systems



Targeted measures to limit display of advertisements



Reinforcing internal processes or supervision



Cooperating with trusted flaggers



Cooperating through Codes of Conduct and Crisis Protocols



European
Commission

Ensuring accountability

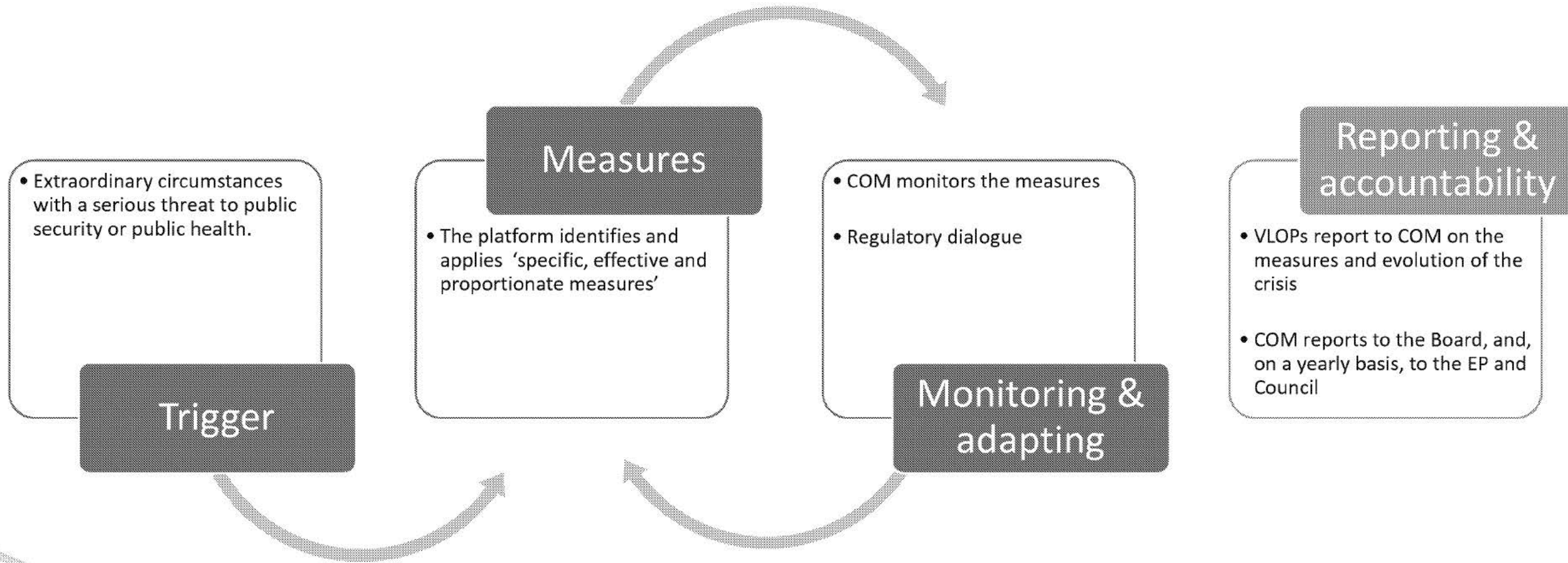
Independent audits

- At least once a year
- Performed by organisations which :
 - Are independent from the very large online platform audited
 - Have proven expertise: risk management, technical competence and capability
- Scope:
 - All due diligence obligations – including risk management measures
 - Commitments taken under Codes of Conduct and Crisis protocols

Public scrutiny

- Public reporting on the risk assessments, mitigation measures, audits
- Specialised scrutiny for the evolution of risks: data access for vetted researchers
- Specialised scrutiny on particular issues: public ad repositories

Crisis response mechanism: fast response & targeted risk management



Recommender systems

Scrutiny & accountability

- Independent audits
- Transparency reporting
- Data access for researchers

Due diligence for very large platforms and search engines

- Risk management obligations

User agency

- All online platforms:
 - meaningful information to users on the general parameters used for recommending content
- Very large online platforms:
 - choice of at least one option which is not based on profiling

User empowerment

Due diligence obligations for platforms

Public scrutiny

Regulatory scrutiny

Advertising

Scrutiny & accountability

- Ads repositories
- Independent audits
- Transparency reporting
- Data access for researchers

Ban on certain types of targeting

Due diligence for very large platforms and search engines

- Risk management obligations

User agency

- Transparency when presenting ads on online platforms: prominent markings, identity of the advertiser, explain why the user is targeted & how parameters can be changed
- Prominent markings for sponsored content (e.g. influencer advertorials)

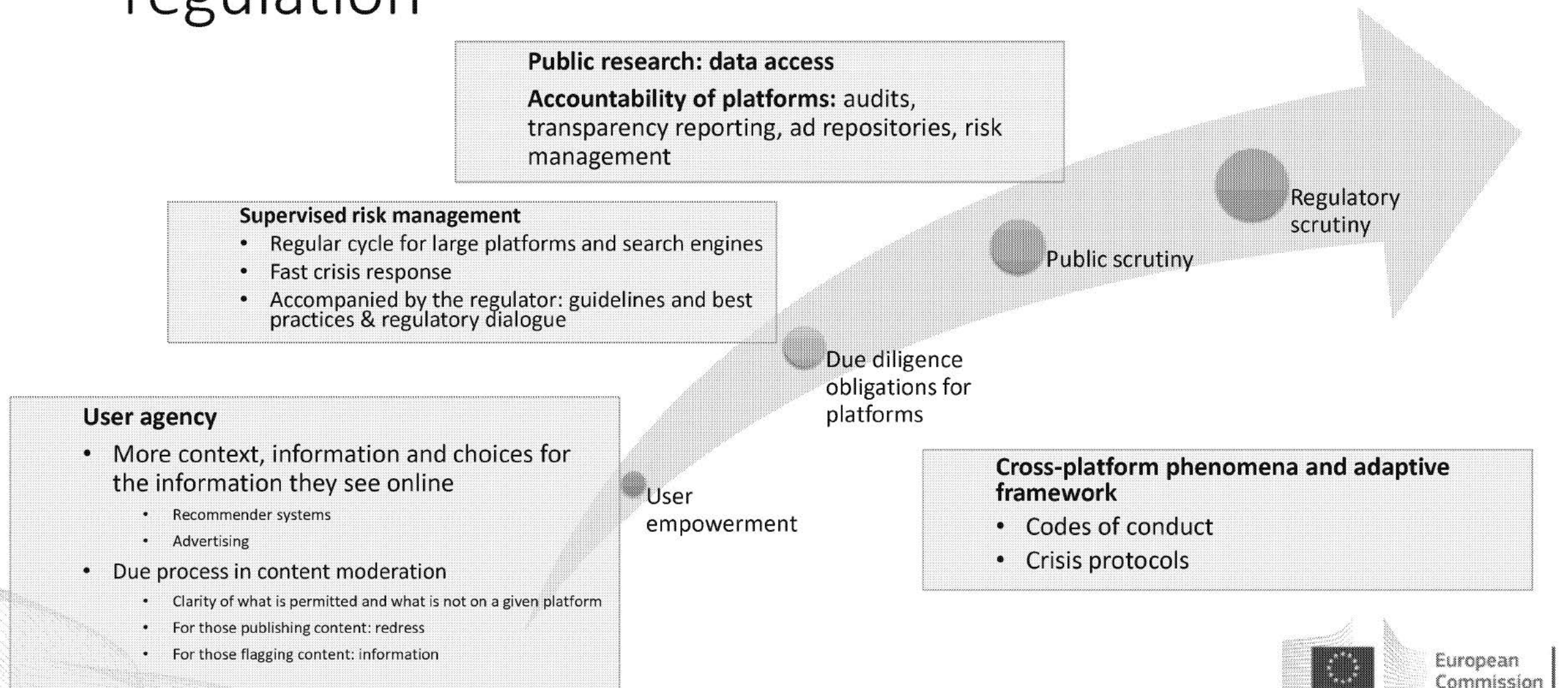
User empowerment

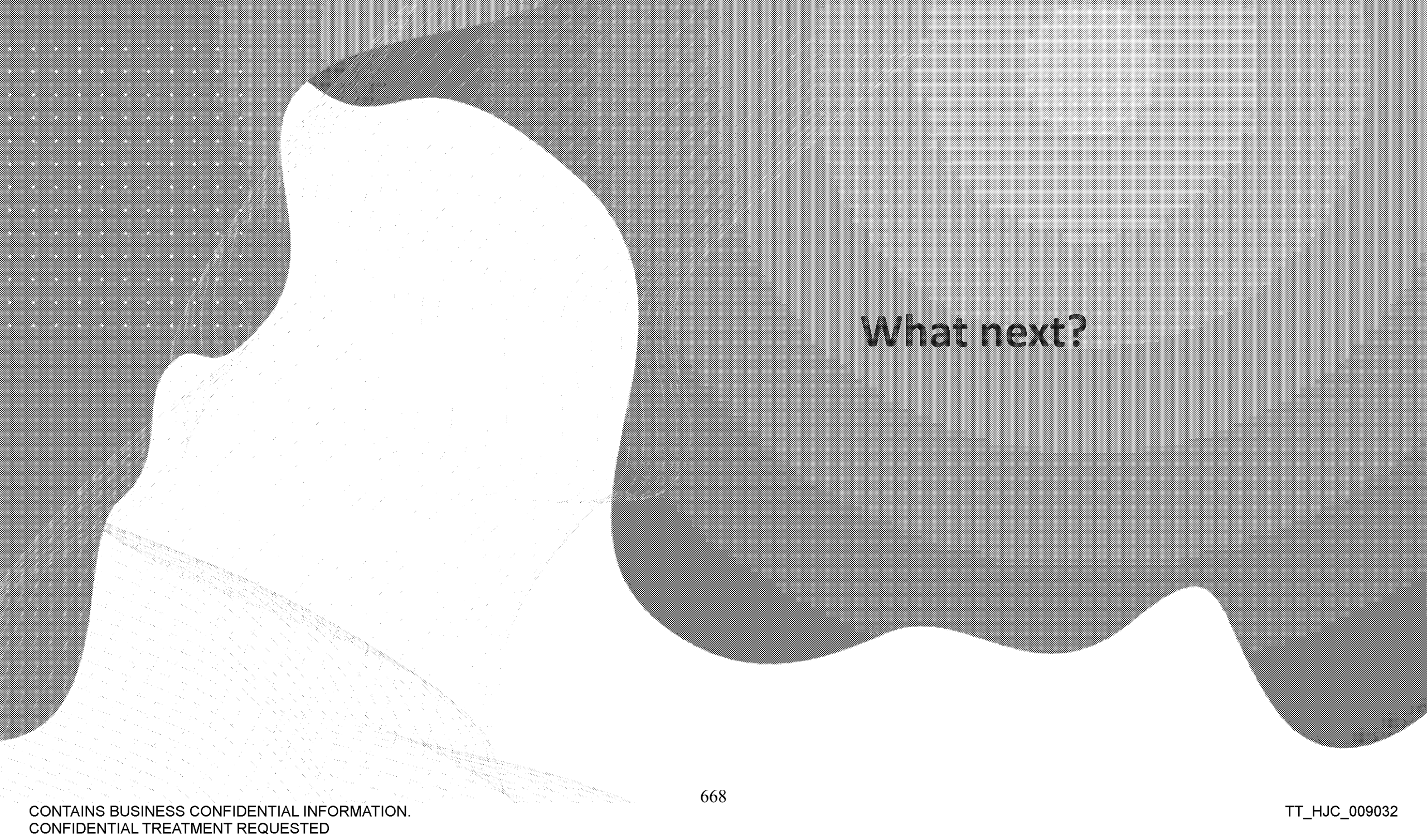
Due diligence obligations for platforms

Public scrutiny

Regulatory scrutiny

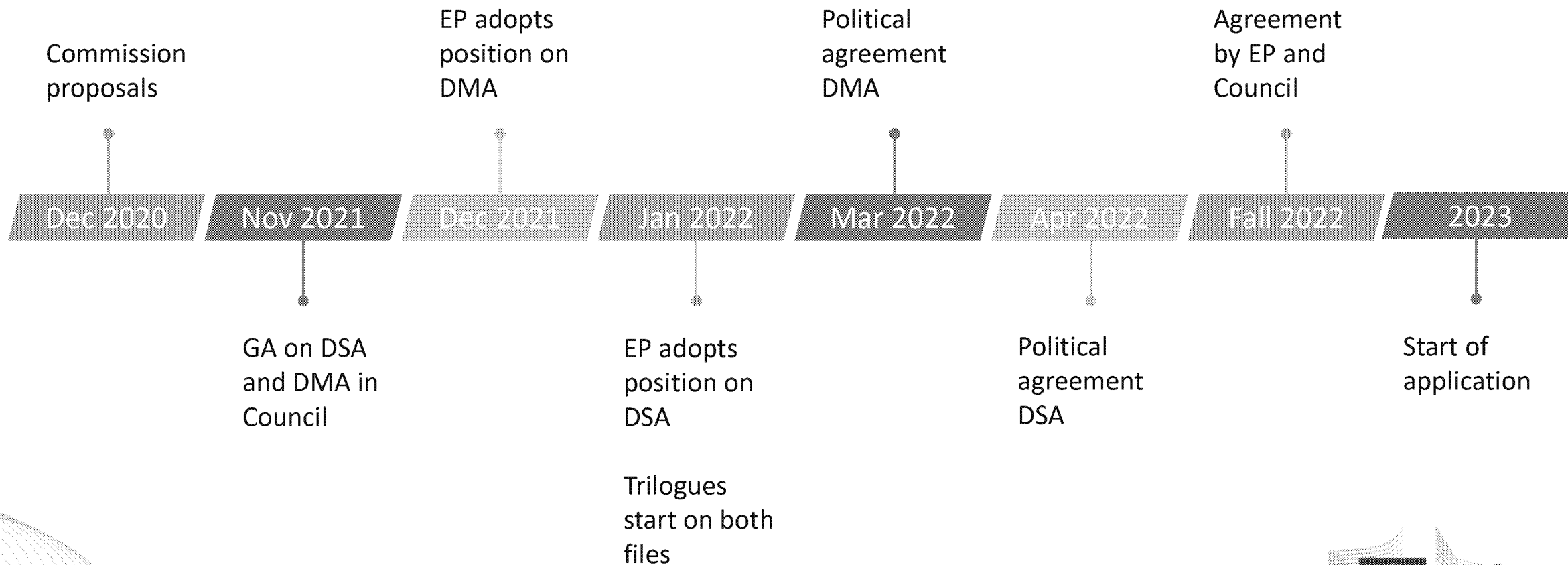
Tackling disinformation: adaptive co-regulation





What next?

Adoption of the DSA & DMA



Our work is only starting

Building capability within the Commission & Member States

Secondary legislation

And we start!



Minister @MikulasBek and @EP_President signed the long-awaited Digital Markets Act 🇸🇰.

The new legislation defines clear rules for large online platforms and will create a fairer space for new players in digital markets.

#EU2022CZ #DMA



Mikulas Bek and 3 others

4:13 PM · Sep 14, 2022 · Twitter Web App

Shaping Europe's digital future

[Home](#) [Policies](#) [Activities](#) [News](#) [Library](#) [Funding](#) [Calendar](#) [Consultations](#)

[Home](#) > [Policies](#) > [European Centre for Algorithmic Transparency](#)

European Centre for Algorithmic Transparency

The European Centre for Algorithmic Transparency (ECAT) is committed to improved understanding and proper regulation of algorithmic systems.

Algorithmic systems determine many aspects of our online experience, for example, a music streaming app may use algorithms to suggest songs or bands to its users. With the ever-increasing societal impact of online platforms such as social networks, online marketplaces, and search engines, there is an urgent need for public oversight of the processes at the core of their business. This includes in particular how these platforms and search engines **moderate content** and how they **curate information** for their users.



Image by montaj - Gettyimages/Block

Thank you!

Redacted

Redacted ec.europa.eu

Redacted



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.



Exhibit 31

Algorithmic Amplification, Borderline Content and Manipulative Techniques

Redacted

29 September 2022, EUIF Workshop

Redacted@isdglobal.org

Redacted

@ISDGlobal

From 'Borderline' to Mainstream: A changing international extremist landscape

- Hypercharged by the Covid pandemic - we have seen the online **threat landscape becoming increasingly hybridised**: the boundaries between disinformation, conspiracy theories, targeted hate, harassment and violent extremism have become ever blurred.
- Mobilisation and threats of violence increasingly come from **a broad church of actors, opportunistically building loose coalitions around crisis points**, shared goals and common objectives.
- Meanwhile, global lockdowns have also catalysed a trend of **'post-organisational' violent extremism**, where the influence or direction of activity by groups or organisations is ambiguous or loose.



Blurred Borderlines: Post-organisational extremism dynamics

- In recent years terrorism and violent extremism across the ideological spectrum have been marked by a **“post-organisational” trend** – where the influence or direction of activity by particular groups or organisations is ambiguous or loose.
- **This is not a new phenomenon.** The notions of “leaderless resistance” and “leaderless jihad” were first discussed decades ago by extremist ideologues such as white-supremacist Louis Beam Jr. and al-Qaeda ideologue Abu Musab al-Suri.
- However **post-organisational dynamics have accelerated with the pandemic**, especially following the mass de-platforming of various actors in January 2021, leading to a new crop of individuals leaving mainstream platforms to join fringe social networking sites where violent extremist ecosystems thrive (Argentino et al)
- Despite the fracturing and franchising of violent extremist movements and the proliferation of decentralised online extremist spaces, responses to terrorist content online are **still hampered by rigid organisational conceptions of the challenge.**

“A confluence of ideological affinities is [becoming] more powerful in inspiring and provoking violence than the hierarchical terrorist organizational structures of the past.”

Redacted

Borderline Content: Conceptualising Post-Organisational Violent Extremism

		Inspirational material Content which can reinforce a violent extremist or terrorist mind-set		Ideological material Material which is specifically trying to further a violent extremist or terrorist world-view		Instructional material Content which contains instructions on operational aspects of terrorist activity	
Violent	Non-Group Violent material not associated with a specific violent extremist group, designed to inspire violent extremism or terrorism, including copycat attacks.	Group Violent material group material designed to inspire violent extremism or terrorism.	Non-Group Material not associated with a specific violent extremist group, which makes the ideological case for extremist violence.	Group Violent extremist group material which makes the ideological case for violence.	Non-Group General instructional material providing operational guidance on carrying out acts of violence and terrorism.	Group Violent extremist group material providing operational guidance on conducting acts of violence and or terrorism.	

- **Post-organisational taxonomy** developed by ISD to serve as a **conceptual framework** for moving beyond solely **group-centred approaches** to understanding violent extremist threats, whilst **ensuring approaches remain robust, transparent and protective of fundamental freedoms**.
- Informed by **analysis of content** shared in **post-organisational violent extremist online spaces**, and **online material** referenced in the **conviction of terrorism offenders** (from Christchurch attacker to ISIS foreign fighters).

Source: A Taxonomy for the Classification of Post-Organizational Violent Extremist and Terrorist Content, ISD for GIFCT, 2021

Borderline Content: Conceptualising Post-Organisational Violent Extremism

	Inspirational material		Ideological material		Instructional material	
	Non-Group	Group	Non-Group	Group	Non-Group	Group
Non-Violent	Material not associated with a specific violent extremist group, which nonetheless inspires the mood music for violent extremism.	Violent extremist group material designed to inspire supporters, but not necessarily to violence.	Material not associated with a specific violent group, which nonetheless helps to build the "system of meaning" underpinning a violent extremist ideology.	Violent extremist group material designed to build the "system of meaning" underpinning a violent extremist ideology.	General instructional material providing operational guidance on non-violent activities relevant to violent extremism.	Violent extremist group instructional material providing operational guidance on non-violent activities.

- **Contextualised approaches** key to moving beyond overly narrow framing of online threats rooted solely in content removal of specific **designated organizations, branded terrorist material, and violent content**.
- **Tiered approaches:** e.g. 'High risk' content flags when non-violent, non-proscribed content circulated by violent extremist communities, such as texts associated with the harmful 'Great Replacement' conspiracy theory.
- **Behaviour-sensitive moderation** taking into account network dynamics: e.g. a user sharing both ideological *and* instructional material could represent a greater risk than someone sharing solely inspirational content.
- Crucially, implementation will require approaches that are **proactively conscious of fundamental rights** and considers policy and product approaches that are more nuanced than the blanket removal of violating content.

Evasion and Amplification: A new set of online violent extremist tactics

Responding to social media company efforts to limit reach of terrorist groups, extremist actors deploy significant efforts to circumvent content moderation of online platforms.

- Content masking
 - overlaying legitimate branding from news sources onto its own images and videos
- Text Disruption/Distortion
 - Broken text or slang phrases can evade automated detection of specific trigger words
- Signposting
 - Sharing links within posts or in comments which lead users to graphic imagery, videos and propaganda stored on file-sharing sites or less restricted platforms
- Co-ordinated Raids
 - Flooding comment sections of posts by public figures or organisations
- Hashtag Hijacking
 - incorporating popular hashtags into posts so they will be seen by unwitting readers
- Account Hijacking
 - Hacking other users' social media accounts to exploit their existing friends or followers and turn the accounts into propaganda hubs

Borderline Content Algorithmic Amplification Case Study: YouTube

- ISD study created 10 accounts presenting as men and boys with a spectrum of right wing ideological interests. We found accounts **tended to be served content based on who they followed**, with 'mainstream' accounts unlikely to be recommended extreme content.
- But regardless of where they were on the ideological spectrum, every single account was **recommended content promoting anti-trans, misogynistic and 'Manosphere' views**.
- Engaging with content featuring Jordan Peterson and Ben Shapiro served as a **gateway** into recommendations for a slew of anti-feminist, misogynistic, and Manosphere content.
- Via this a 'blank account' was recommended **fashwave** aesthetic content, music edits of **Nazi soldiers, Hitler fan videos** and docs about **serial killers** with female victims.
- Qualitatively shows the **significant overlap** between **Manosphere** and **white supremacist online subcultures**, and the role algorithms can play in facilitating user journeys.

Redacted

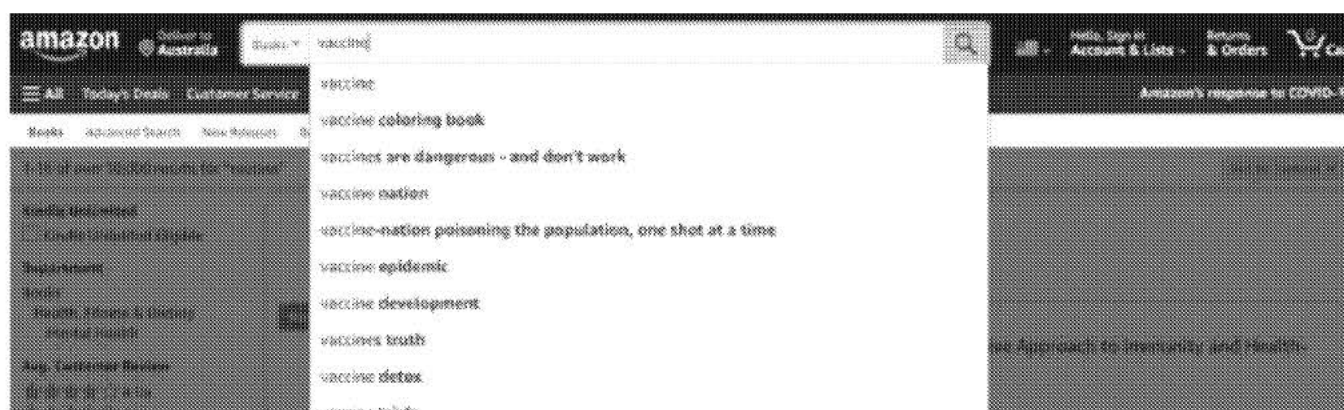
Algorithmic Amplification Beyond Social Media: Amazon

ISD research on **Amazon's book sales platform** illustrates how problems with algorithmic recommendation of borderline content **extend beyond social media platforms**.

ISD research showed the platform was promoting dangerous conspiracist literature to users searching for book titles on their platform, serving to **cross-propagate conspiracy theories**, **recommend hardline** ideological material, and **auto-populate harmful searches**.

At the core is an apparent failure to consider risks for when what a system designed to upsell customers on fitness equipment or gardening tools is **unleashed on products espousing conspiracy theories, disinformation or extremism**.

The question of **banning books is highly contentious**. But the problem of algorithmic amplification could be mitigated by turning recommendations off on such products, avoiding actively promoting harmful content and funnelling more money into the pockets of creators.



Source: Recommended Reading: Amazon's algorithms, conspiracy theories and extremist literature, ISD, 2021

Implications for policy and practice

The borderline is shifting – with the threat less explicitly ‘group’ based and increasingly intersecting with other information harms like conspiracy theories. Extremists are also adopting an ever more transnational perspective. This requiring a new generation of responses, which go beyond existing counter-extremism policy paradigms. This includes:

- **Regulation:** Consider human-rights based approaches to social media regulation which focus on systemic risk mitigation and transparency rather than content removal alone, recognising platform design issues which facilitate the growth of extremist movements.
- **Intervention:** Develop the next generation of offline and online interventions to address these looser movements and bringing people out of extremist ecosystems.
- **Prevention:** Formulate programmes, including curricula and community initiatives, required to raise awareness of and build resilience against hybrid extremism threats. Greater efforts should be made to educate publics and develop critical thinking, including partnering with the private sector, schools and universities, religious and youth organisations.
- **Coordination:** Facilitate improved regional policy exchange, recognising the transnational nature of these threats, and consider lessons learned from mechanisms developed for international collaboration to counter specific groups (e.g. Counter Daesh Coalition).

Thank you.

Redacted@sdglobal.org

Exhibit 32

From: HOME-INTERNET-FORUM@ec.europa.eu [HOME-INTERNET-FORUM@ec.europa.eu]
Sent: 10/11/2022 1:20:23 PM
To: HOME-INTERNET-FORUM@ec.europa.eu
CC: [Redacted]@ec.europa.eu; [Redacted]
[Redacted]@ec.europa.eu; [Redacted]@ec.europa.eu;
[Redacted]@ec.europa.eu; [Redacted]@ec.europa.eu;
[Redacted]@ec.europa.eu; HOME-NOTIFICATIONS-D3@ec.europa.eu
Subject: [External] FLASH REPORT: EUIF Workshop on Algorithmic Amplification and Borderline Content - 29 September 2022
Attachments: Presentatie EU Internetforum def.pdf; EUIF - EOOH .pptx (public).pdf; DSA EUIF 29 Sept 2022.pdf; EUIF presentation_ISD.pdf; 20220929_EUIF_presentation_PEReN_v2.pdf; Romanian Intelligence Service - Presentation for EUIF Workshop on borderline content (Brussels 29.09.2022).pdf; GIFCT EUIF Algorithms.pdf; Trust Lab Presentation 9_29.pdf

Dear participants to the EUIF workshop on algorithmic amplification and borderline content,

Thank you very much for joining us in person and online and for contributing to a lively and fruitful discussion. Please, find hereby a flash report of the workshop and some of the PPT presentations in attachment.

We look forward to seeing/e-seeing you in the upcoming EUIF events.

Kind regards
The EUIF Team

FLASH REPORT: EUIF Workshop on Algorithmic Amplification and Borderline Content **29 September 2022**

Summary

On 29 September the Commission organised an EUIF workshop to explore the possible negative effects of algorithmic amplification techniques on the user journey towards radicalisation and provide guidance to tech companies on possible definitions and thresholds concerning the spread of borderline content.

➤ **Main takeaways**

- Examples and description of borderline content shared by Member States, EU services and researchers reveal a **combination of disinformation/conspiracy theories and hate speech**.
- Most common **type of content** identified by participants anti-establishment/anti-institutions, anti-Jewish, anti-trans, misogynistic, anti-migrants, racist, anti-COVID measures.
- Both **safety by design measures** and **positive interventions** need to be adopted in order to prevent the spread of borderline content via automated systems.
- Focus on content is not enough. Preventive measures must take into account **behavioural patterns and propagation tactics** used by malicious actors, including **manipulation**, and include digital and media literacy, critical thinking and democracy-strengthening programs to foster resilience.
- **Definitions and taxonomies** are essential basics to build effective preventive measures.
- Other suggested technical solutions were: **downranking**, **demonetise**, add **friction** to access.
- As regards government's lists for content moderation, companies suggested they are not enough, as they may be politically biased and incomplete. **Civil Society Organisations** should be empowered to compile lists and build taxonomies.
- Challenges in connection with the use of **non-cooperative platforms** and **Terrorist Operated Websites** to spread TVEC and borderline content leading to violence were discussed.

➤ **Next Steps**

- Outputs of the meeting will be gathered in a **handbook/brochure to be used** by tech companies as voluntary-based guidance on content moderation of borderline content
- The deliverable will be built around content falling under the definition of **hate speech supported by conspiracy theories** and disinformation, which can lead towards violent extremist behavior.
- As agreed in the workshop, the ongoing **EUIF study on algorithmic amplification** will benefit from the contribution of experts involved in the workshop and from companies
Preliminary results of the study will be shared at the beginning of 2023 and final results in Spring 2023.
- Upcoming EUIF meetings:
 - o Technical Meeting on Violent Extremist and Terrorist Financing Activities Online on 12 October
 - o Senior Official meeting on 16 November,
 - o Ministerial meeting on 7 December

In details

State of play of actions taken by the EUIF and its members

- **EUIF study** on the effects of the misuse of algorithmic amplification to spread TVEC and borderline content on the user journey has been launched. The **methodology** was presented. The study raised interest by Member States.
- GIFCT updated about latest research by GNET of role of **users agency** in the functioning of recommender systems, the asset of the new GIFCT working groups for 2022 and the state of play of work on definitions and designations
- TAT raised the importance of dedicating more attention to the **role played by TOWs** in the dissemination of TVEC and other content that may be more difficult to identify
- **Techniques used for the propagation** of the content is another important element to analyse further, as well as the positive effects of **designation** as a way to legitimate and foster content moderation
- Microsoft referred to the **Christchurch Call Initiative on Algorithmic Outcomes**, which is undertaking together with the U.S., New Zealand and Twitter. It will explore how privacy-enhancing technologies can help drive greater understanding of algorithmic outcomes. This pilot project will develop new software tools that can help facilitate more independent research on the impacts of user interactions with algorithmic systems.
- Meta is working on a strategic network disruption protocol targeting abusive behavior and keep track of tactics used, with features that prevent reinstatement of malicious users on the platform
- Google/YouTube has been working on definitions and thresholds regarding the use of borderline content to ensure de-ranking by its internal algorithmic systems

Presentations by DG CNECT , DG JUST and EEAS

- DG CNECT provided a complete overview of the provisions of the Digital Services Act concerning algorithmic transparency, obligations to carry out risks assessments and how the act is addressing disinformation

Other policy initiatives addressing the spread of hate speech and disinformation:

- Cooperation and empowering civil society, e.g. through **the European Digital Media Observatory - EDMO**
- Cooperation with social media platforms: Code of Practice against Disinformation, Code of Conduct on Illegal Hate Speech.
- The Creation of the **European Observatory of Hate Speech (EOOH)** by DG JUST
- EEAS new instrument: **Foreign Information Manipulation and Interference (FIMI)** –shifting from the mere focus on content to looking at the behaviour of actors (foreign actors is the focus for EEAS), i.e. the manipulative tactics, techniques and procedures that are being used to distort public opinion and civic discourse

Member States' perspectives

Presentations were given by Netherlands, Italy, Spain, Romania and France (while the presentation prepared by Czech Republic could not be given due to technical problems)

Key takeaways from Member States presentations

- Examples of content shared by Member States included **posts, pictures and memes**, including humorous ones, targeting the Jewish community, migrants, feminism and reference to conspiracy theories on COVID crisis and the war against Ukraine
- A number of Member States referred to **anti-establishment/elites movements** and their use of platforms to incite violence during anti-COVID and anti-VAX protests – shift from COVID to pro-Russia narratives were also observed
- Reference to the Commission's **PBC on anti-vax/anti-government movements** to underline connections between new emerging forms of extremism and the growing use of borderline content
- The use of **tactics to propagate borderline content** were also explained, such as the use of influencers raising not only reactions, but also a high number of views and use of algorithmic amplification techniques,
- Member States also referred to the use of **cryptocurrency** linked to borderline content to avoid detection
- **Lack of groups designations** is also a challenge for moderation and prevention of violence
- Two Member States presented technical means to detect content and carry out **cross-platforms analysis**
- Possible technical solutions and methodologies to **audit algorithmic amplification** systems and analyse propagation of borderline content without disclosing personal data were also presented

Research/studies presented by ISD, Moonshot and UNOCT

- The Institute for Strategic Dialogue (ISD) presented a taxonomy of **post-organisational violent extremism**, and use of ideological, inspirational and instructional material
- The presentation highlighted a major change in the online threat picture, due to **new forms of extremism**, described as “post-ideological”
- ISD identified the misuse of **algorithmic amplification as gateway** to more extremist content
- Moonshot presented a study that coded content to cluster it between violent extremist and borderline and confirmed that **behavioural patterns** affected the ranking
- **Opportunistic attitudes** and fluidity were observed among different types of hate speech
- Moonshot has been involved in **positive intervention initiatives**, such as the re-direct method to divert users' attention towards positive narratives
- Such tactics works but must be combined with preventive measures, especially security by design
- UNOCT presented a Research Launch on *Examining the Intersection Between Gaming and Extremism* addressing **dissemination of borderline content on gaming platforms** also through misuse of **algorithmic amplification** techniques

Useful links shared during the meeting

On the **European Observatory of Online Hate** (DG JUST funded): <https://eooh.ai/>

On **Microsoft** new Initiative on algorithmic amplification: <https://blogs.microsoft.com/on-the-issues/2022/09/20/christchurch-call-responsible-ai-online-extremism/>

On **Meta's** approach to research and transparency FORT to facilitates data sharing and the publication of independent research about Facebook's role in society, with the right privacy protections in place: <https://fort.fb.com/>

On **YouTube's** cooperation with researchers <https://research.youtube/>

UNOCT shared a link to the event to launch their study on videogaming [Bit.ly/3qFmVnc](https://bit.ly/3qFmVnc)

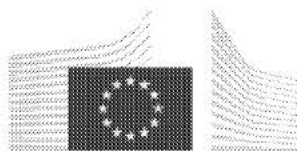
Redacted

FROM GIFCT

- The Contextuality of Lone Wolf Algorithms: An Examination of (Non)Violent Extremism in the Cyber-Physical Space: <https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-ContextualityIntros-1.1.pdf>
- Methodologies to Evaluate Content Sharing Algorithms & Processes GIFCT Technical Approaches Working Group: <https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TA-Evaluate-1.1.pdf>
- Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence GIFCT Transparency Working Group: <https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf>

- Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes: <https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAP11-2021.pdf>
- A range of GNET Insights that touch on the topic of Algorithmic Amplification and extremism: <https://gnet-research.org/tag/algorithm/>

Exhibit 33



EUROPEAN COMMISSION

DRAFT AGENDA
EU INTERNET FORUM
MINISTERIAL MEETING
hosted by
COMMISSIONER [Redacted]

7 DECEMBER 2022
14:30 - 17:45 (CET)

HYBRID MEETING
IN PERSON VENUE : BERLAYMONT BUILDING (SCHUMAN ROOM), BRUSSELS

INTERPRETATION:
EN-FR-DE-ES-IT-CZ

14:15-14:30 *Family photo (for physical attendees)*

14:30-14:50 **Opening remarks (public session)**

- **[Redacted]** *European Commissioner for Home Affairs*
- **[Redacted]** *Minister of Interior of Czech Republic, EU Council Presidency*

14:50-16:10 **First debate: *Preventing and Combating Child Sexual Abuse and exploitation online***

- How can technologies detect child sexual abuse online in unencrypted and encrypted systems?
 - *Introduction by [Redacted] staff data scientist at Thorn*
 - *Presentations by companies*
- What is required for effective reporting to support identification of victims and investigations of child sexual abuse?
 - *Introduction by [Redacted] European Multidisciplinary Platform Against Criminal Threats (EMPACT)*

16:10-17:30 Second debate : *Countering violent extremism and terrorism online*

- What is the current threat landscape, modus operandi and how terrorists and violent extremists misuse new technologies?
 - *Introduction by Professor [Redacted] King's College, London*
- How is the industry addressing borderline content and enhancing transparency of recommender systems?
 - *Presentations by companies*
- What are the lessons learnt from the Bratislava attack?

17:30-17:45 Closing remarks

- [Redacted], *European Commissioner for Home Affairs*

Exhibit 34

From: [Redacted]@ec.europa.eu]
Sent: 1/9/2023 5:14:01 PM
To: ve_home.d.3 (HOME) [home-internet-forum@ec.europa.eu]
CC: [Redacted]@ec.europa.eu]; [Redacted]
[Redacted]@ec.europa.eu]; [Redacted]
[Redacted]@ec.europa.eu]; [Redacted]@ec.europa.eu]; [Redacted]
[Redacted]@ec.europa.eu]; [Redacted]@ec.europa.eu]; [Redacted]
Subject: [External] FLASH REPORT: EU INTERNET FORUM (EUIF) MINISTERIAL MEETING - 7 December 2022 - Ares(2023)138069
Attachments: image001.png

FLASH REPORT: EU INTERNET FORUM (EUIF) MINISTERIAL MEETING - 7 December 2022 - Ares(2023)138069 (Please use this link only if you are an Ares user – Svp, utilisez ce lien exclusivement si vous êtes un(e) utilisateur d'Ares)

Dear EUIF Members,

Thank you for participating in the EUIF Ministerial meeting that took place in hybrid form on 7 December 2022. Please, find below a flash report of the meeting.

Kind Regards,

The EUIF team



European Commission
Directorate-General Migration and Home Affairs
Directorate D: Internal Security
Unit D3: Prevention of Radicalisation

E: HOME-INTERNET-FORUM@ec.europa.eu

FLASH REPORT: EU INTERNET FORUM (EUIF) MINISTERIAL MEETING **7 December 2022**

[Redacted] hosted the 8th EUIF Ministerial meeting, which took place in hybrid format, in Brussels and online.

Summary

The objective of the meeting was to address the **shared responsibility of stakeholders to ensure security online**. Various speakers and presenters provided insights on new trends concerning criminal activities, the importance of aligning children's rights with the digital environment and content moderation of terrorist, violent extremist and borderline content online.

It was clear from the interventions that proactive **outreach from tech companies to law enforcement agencies (LEAs)** is crucial. Participants agreed on the added value of the EUIF to facilitate exchanges and the need for

further collaboration with industry to deliver on the shared responsibility of making the online space safe, especially for children.

The **first topical debate** focused on the **fight against child sexual abuse (CSA)** and on **keeping children safe online**. Participants who took the floor reiterated their commitment to this effort and expressed support for the legislation to combat and prevent CSA, noting specifically the need for **clear obligations** on companies to detect, report and remove child sexual abuse online. A spotlight was drawn on **possible tools to detect known and new (previously unseen) child sexual abuse material (CSAM)** and **grooming** in unencrypted systems, and also their application in encrypted systems as well as **reporting**, and the important role it plays to help identify and safeguard children from abuse and disrupt sharing of CSAM. Participants discussed the importance of having detection mechanisms embedded also in a broader and encompassing response to CSA, that includes robust prevention mechanisms. Participants outlined the need for robust and standardised mechanisms and templates to report CSA and the specific critical information – i.e. the illegal content - that is required to make a report actionable.

The **second topical debate** focused on **emerging challenges when countering violent extremism and terrorism online**. Participants agreed that violent extremist and borderline content is now **proliferating on smaller platforms**, which have limited resources or willingness to effectively moderate this content. Moreover, participants underlined the **importance of addressing challenges posed by the increasing use of borderline content**, including forms of misogynist, discriminating and dehumanising content, also known as harmful but legal content that comes close to infringing on the community guidelines of platforms or laws regulating online illegal content. Participants also **discussed ways in which the industry should enhance transparency** on the criteria followed by platforms to recommend targeted content to users through **algorithms**. The online dimension of the Bratislava terrorist attack was also discussed as an example of how the decentralisation of content dissemination can pose major difficulties to content moderation efforts by platforms and law enforcement agencies.

Details

In her opening remarks Redacted spoke about the scourge of **child sexual abuse**. She highlighted the crucial role online service providers play in ensuring children are safe on their services and thanked them for their ongoing work to detect, report and remove child sexual abuse. She touched upon the ongoing negotiations on the proposal for a Regulation to prevent and combat child sexual abuse, noting the fundamental role of prevention in the proposal. The Commissioner called for support to reach an effective agreement on the proposal before summer of 2024. If this is not achieved, companies will no longer be able to take effective action to prevent, detect and report child sexual abuse. The proposal provides a balanced approach that safeguards all fundamental rights at stake.

On **terrorism and violent extremism online**, the Commissioner referred to recent attacks in Bratislava and Sweden, which resulted from the spread of online violent right wing extremist content targeting, among others, the Jewish and LGBTIQ communities and women. She expressed concerns for the increased presence of borderline content online and the challenges it brings about when it comes to content moderation. She also warned against the growing use of terrorist operated websites, manipulation techniques used by extremists and the need for more transparency on the threats related to the misuse of algorithmic amplification techniques. EU regulations concerning the online space will mitigate part of the challenges ahead, but voluntary cooperation is crucial and must continue.

Redacted underlined the importance of the work done in the EU Internet Forum and affirmed that Child Sex Abuse is a priority for Czech presidency. He highlighted the progress made on the

negotiations under the Czech presidency, including organising two technical workshop, one covering detection technologies and the other looking at age verification mechanisms. [Redacted] stressed the need to put pressure on MSs to focus on this issue and to implement the new regulation, and committed Czechia to continue supporting the discussions in the Council. He also referred to a new wave of propaganda and disinformation narratives in Europe, as well a rise of anti-Semitic and anti-Democratic movements, and that the terrorist attack in Bratislava was one of the consequences of the rising of these movements. Finally, [Redacted] called up on EU Internet Forum to deal with borderline content and help come up with ways to prevent the spread of this content.

1. Detect child sexual abuse online

The Commission emphasised the urgent need to adopt the Regulation to ensure that companies can do their part to prevent child sexual abuse, through measures that focus on child safety by design, also beyond the expiry of the interim derogation. Where necessary, there is a possibility to mandate companies to detect, report, block and remove child sexual abuse online.

[Redacted] initiated the discussion with a presentation on existing technologies to detect known and new or previously unseen child sexual abuse material and grooming in unencrypted systems, and also their application in encrypted systems. He provided a high-level explanation on the technical functioning of the different detection solutions and listed examples in practice, including information on Thorn's research efforts and existing services.

His presentation was followed by interventions from company representatives on specific technologies utilised on their platforms and services. Companies elaborated on specificities related to detection of CSA in their services, the tools they make use of and the existing challenges. Participants agreed on the importance of comprehensive approaches that embed the detection of CSA also in a broader setting that is built on robust architecture guided by safety-by-design measures to prevent child sexual abuse from happening in the first place. A number of participating companies also commented on the need for enhanced education and awareness raising for child users and their caregivers on avoiding risky situations online and empowering children in particular to report to companies bad/suspect behaviours.

Participants recognised that reporting these abuses in a manner that allows for identification of victims and accountability of perpetrators is instrumental. [Redacted]

[Redacted] explained what is required for effective reporting to support identification of victims and investigations of child sexual abuse, notably the fact that meta data alone without the content were insufficient to launch the investigations. **EU Member States** agreed on the necessity and importance of fighting child sexual abuse online, and highlighted the importance of the proposed regulation and of clear obligations of service providers. Additionally, they pointed to the need for clear and consistent regulation as well as the importance of using AI to prevent such behaviours, and emphasised the importance of launching effective and coordinated investigations, which should remain a priority in the next years. There is a fundamental need for joint efforts, particularly in cooperation and communication, and it is essential to involve companies in all work to create good practices, asking for more EU funds for these projects. A number of participants also mentioned the importance of strengthening cooperation on this topic with neighbouring regions, such as the Western Balkans, to help build knowledge and capabilities of governments and competent national authorities including law enforcement, to set out an effective and global response to these crimes.

Transparency reporting accompanying companies' efforts is another key consideration to raise accountability and ensure efficacy of measures taken by companies in this space. It also serves to share good practices and promotes better collaboration.

The Czech Republic informed about two workshops on technology and age verification under its Presidency, and Sweden confirmed that this will be a priority in their upcoming presidency, considering to include in the agenda the need for high-quality reporting to fight child sexual abuse.

2. Countering violent extremism and terrorism online

Redacted introduced the session focusing on the online threat landscape and the challenges ahead, with particular reference to the increasing presence and spread of violent right wing extremist content online ever since the Christchurch attack in 2019. He highlighted three challenges: **1)** The merging of online spaces that were previously used to spread conspiracy theories related to the COVID-19 pandemic that are now used to disseminate pro-Kremlin disinformation aimed at sowing mistrust in our societies and radicalise users mostly in Telegram **2)** Increase in gender-based violence (which can include misogynist content/views that promote violence against women, such as within incels communities online and against the LGBTIQ+ community); **3)** Effects of Twitter's decision of reinstating extremist accounts that were previously banned for promoting pro-nazi and fascist propaganda.

Redacted added that in the next year we will still have long-standing trends of domestic and transnational terrorism online and offline. She acknowledged that nowadays we face a 'Pick and mix' of ideologies, as some extremists are not part of a specific organisation nor ideology. This new landscape brings up the need to focus on different kinds of online behaviour and content. Extremists no longer share only videos and photos, but also use PDFs, URLs, and audios to disseminate their manifestos (e.g the Bratislava manifesto). Countries face now a challenge on how to deal with borderline content, and that is the reason why there is a need to create handbooks and other materials that can help to provide definitions of grey areas. GIFCT informed they will provide clarity around these buckets. Representatives from the **industry** affirmed they have a zero tolerance approach regarding online extremism and reported about measures taken to prevent use of algorithmic amplification to spread not only TVEC but also borderline content. There was also reference to the joint collaboration between Microsoft, Twitter, NZ and USA in giving access to researchers to better understand algorithmic outcomes. Companies also presented their approaches to the moderation of harmful content that goes beyond requirements established by laws and that aims at keeping users safe from hateful and dehumanising content and reported on efforts to prevent terrorists and violent extremists from using virtual reality and new technologies to create immersive content that can lead towards radicalisation.

3. Lessons Learnt from the Bratislava terrorist attack

Europol reported on the spread of content shared by the perpetrator across different platforms after the attack and called for law enforcement agencies to update their detection capability and ensure swift collaboration with tech companies on data preservation and disclosure of information.

The EU Counter Terrorist Coordinator outlined concerns about the spread of hate speech and borderline content and mentioned the progress made by establishing EU Regulations, such as the Terrorist Content Online regulation and the Digital Services Act, and the importance of ensuring their smooth implementation as strong preventive measures to the spread of illegal and harmful content. He also addressed the importance of increasing algorithmic transparency and improve coordination between platforms to avoid cross-platform dissemination of violent extremist content. Finally he pointed at the migration of extremist groups towards alternative tech platforms with less content moderation, strongly encouraging all the companies and GIFCT to engage with these platforms.

A number of Member States expressed their concern reading the algorithmic amplification and called for more cooperation between authorities and companies. Member States also confirmed that borderline content represent a big concern and it should be a priority and should be addressed jointly. It was also underlined by Member States that extremists' strategies are shifting quickly and a fast approach is therefore needed, as well

as the urgent need to get ready for the metaverse and immersive technologies (*any technology that extends reality or creates a new reality by leveraging the 360 space*). Member States also highlighted the importance of safeguarding freedom of speech when dealing with borderline content.

All the Members States welcomed the work done by the EU Internet Forum in providing instruments and guidelines for the prevention of online violent extremism and terrorism such as the Knowledge Package on violent right wing groups, symbols and manifestos, video-gaming and the Handbook on borderline content.

The Under-Secretary-General at United Nations Office of Counter-Terrorism noted that UN is finalising the Global Terrorism Strategy Report for the 8th review next June and called for Member States to cooperate with the UN in finding joint approaches and solutions.

Conclusion

Redacted concluded by thanking all the participants for their interventions and engagement that allowed to go deeper on some of the challenges ahead and encouraged to continue joint efforts and collaboration. She also welcomed that the EUIF has also addressed this year other threats online and stressed that this work should continue.

EU Internet Forum: Unified action needed in the fight against child abuse online and terrorist content online (europa.eu)

EUIF brochure_en.pdf (europa.eu)

Exhibit 35

From: [Redacted]@ec.europa.eu]
Sent: 9/2/2024 1:47:38 PM
To: [Redacted]@tiktok.com]; HOME-NOTIFICATIONS-D3@ec.europa.eu
Subject: RE: [External] FW: Invitation - EUIF informal focus group on algorithmic amplification
Attachments: Concept Note_EUIF technical meeting on TVE financing activities online.pdf; Agenda_EUIF technical meeting on TVE Financing Activities Online.pdf

Dear [Redacted]

Excellent, I will forward the email with more information to [Redacted] (with you in cc). Could you let me know if [Redacted] will join us in person or remotely? I am looping in my colleague from @HOME NOTIFICATIONS D3 to follow-up accordingly.

On another event, will you or a colleague be able to join us for the EUIF technical meeting on financing activities the day before (10 September)? We had a call on this with another colleague from your side before the summer and I think it would be greatly appreciated to have TikTok participate. I re-attach the invitation and would welcome registration at your earliest convenience.

Best regards,

[Redacted]

From: [Redacted]@tiktok.com>
Sent: Monday, September 2, 2024 3:07 PM
To: [Redacted]@ec.europa.eu>
Subject: Re: [External] FW: Invitation - EUIF informal focus group on algorithmic amplification

And I just got a name - please send the event information to [Redacted]@tiktok.com. Grateful if you could Cc me in on any communication as well.

Thank you,

[Redacted]

From: [Redacted]@ec.europa.eu>
Date: Mon, Sep 2, 2024, 2:25 PM
Subject: RE: [External] FW: Invitation - EUIF informal focus group on algorithmic amplification
To: [Redacted]@tiktok.com>

Dear [Redacted]

Thank you for your prompt reply despite all. It is great to hear that TikTok will join. I will continue sending the information to you until I receive the name of your colleague.

Best regards,

[Redacted]

From: [Redacted] <[Redacted]@tiktok.com>
Sent: Monday, September 2, 2024 2:13 PM
To: [Redacted] <[Redacted]@ec.europa.eu>
Subject: Re: [External] FW: Invitation - EUIF informal focus group on algorithmic amplification

Thank you for your kind words, it is appreciated.

We will attend the focus group but I still need to confirm our representative. As soon as I have a name, I will share with you. I am unfortunately not available on that date.

Kind regards

[Redacted]

From: [Redacted] <[Redacted]@ec.europa.eu>

Date: Mon, Sep 2, 2024, 10:57 AM

Subject: [External] FW: Invitation - EUIF informal focus group on algorithmic amplification

To: "[Redacted] <[Redacted]@tiktok.com>" <[Redacted]@tiktok.com>

Dear [Redacted]

I am deeply sorry to have read about the passing away of your brother. My thoughts are with you and your family and please accept my sincere condolences..

I can hardly imagine how you must feel and how difficult it must be for you to focus on work in this current context. In this, I am sorry to email with a request so shortly after your return.

Following the summer break on top, I am sure you have a lot to catch-up on. To avoid that the below falls to the cracks, I kindly wanted to remind of the invitation to the informal focus group to start the work on the set of principles on algorithmic amplification. Will TikTok be able to join us?

As mentioned earlier, especially with your latest changes to policies regarding visibility in the ForYouFeed, I think the participation of TikTok would be really valuable.

Please feel free to refer me to a colleague and/or get back at your speed.

I am wishing you and your family a lot of strength during this time

Best wishes,

Redacted

From: HOME INTERNET FORUM <HOME-INTERNET-FORUM@ec.europa.eu>
Sent: Wednesday, August 21, 2024 4:50 PM
To: HOME INTERNET FORUM <HOME-INTERNET-FORUM@ec.europa.eu>
Cc: **Redacted** @ec.europa.eu
Subject: Invitation - EUIF informal focus group on algorithmic amplification

Dear EUIF member,

We hope this email finds you well rested after the summer.

You are receiving this email because you have previously expressed interested in the EUIF work on algorithmic amplification.

WHAT is planned? In the EUIF activities 2024, the EUIF announces to develop a set of principles for companies to address algorithmic amplification, including of borderline content on their platforms. We would like to commence this work with a small, informal 'focus group' of interested MS and companies to ensure buy-in and expertise from Forum's members and ensure that companies are willing and capable of adhering to this voluntary set of principles.

HOW will we do it? The idea is to set up an informal 'focus group' which will develop a first draft set of principles which will then be shared with wider EUIF members for feedback. In an initial meeting, the focus group can exchange on key challenges and gather first ideas of which principles, steps and commitments are considered necessary and feasible to address the issue of algorithmic amplification which can contribute to the user's journey towards radicalisation. Based on this initial meeting, we will either convene a second informal meeting or circulate a draft document to continue fine tuning the set of principles before it is shared with all EUIF members for feedback.

WHEN & WHERE will we meet? 11 September, at 10:30 am, hybrid (Brussels). We take the opportunity to invite all attendees of the EUIF technical meeting on 10 September to attend this informal focus group meeting on 11 September in person. For those unable to travel to Brussels, we will provide a remote connection.

WHO should attend? We aim to bring together EUIF Member States and companies to have an active exchange of ideas and build a robust set of principles. For you to assess who should participate in the meeting -but without pre-empting the work of the focus group- it is likely that we will develop high-level, political principles rather than diving into technical details. If you are joined by your engineers or technical experts or come equipped with an understanding of what is possible and what is harder to implement from a technical perspective, it would certainly be beneficial for the process.

Please reply to this email and confirming or declining participation in the small focus group by 2 September, and if you are able to participate in person or remotely.

We are looking forward to hearing from you.

Kind regards,

The EUIF team



European Commission

Directorate-General Migration and Home Affairs

Directorate D: Internal Security

Unit D3: Prevention of Radicalisation

E: HOME-INTERNET-FORUM@ec.europa.eu

Exhibit 36

From: [Redacted]@tiktok.com]
Sent: 1/17/2025 8:48:35 AM
To: HOME-INTERNET-FORUM@ec.europa.eu
Subject: Re: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024 - Ares(2024)7942846
Attachments: EUIF Questionnaire - Algorithmic Amplification_TT response.pdf

Good morning all,
I hope you're well.
With apologies for the delay, sharing our feedback on the questionnaire, in attachment.
All the best,

[Redacted]

[Redacted] Brussels
[Redacted]
[Redacted]
EU Transparency Register: [Redacted]
[Redacted]

From: HOME-INTERNET-FORUM@ec.europa.eu
Date: Thu, Jan 16, 2025, 9:59 AM
Subject: RE: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024 - Ares(2024)7942846
To: "[Redacted]@tiktok.com">

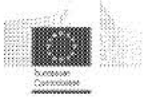
Dear [Redacted]

Happy new year to you too!

By the end of this week is still okay, thank you for the information.

Kind regards,

The EUIF team



European Commission

Directorate-General Migration and Home Affairs

Directorate D: Internal Security

Unit D3: Prevention of Radicalisation

E: HOME-INTERNET-FORUM@ec.europa.eu

From: A [Redacted] <[Redacted]@tiktok.com>

Sent: Thursday, January 16, 2025 9:47 AM

To: HOME INTERNET FORUM <HOME-INTERNET-FORUM@ec.europa.eu>

Subject: Re: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024 - Ares(2024)7942846

Dear all,

I hope you're well, and that you had a good winter break.

I'm very sorry for the delay in getting the questionnaire back to you, I am just waiting on some final feedback and should be able to share it by the end of the week, if this is ok?

Kind Regards,

[Redacted]

--

Redacted

Brussels

Redacted

Redacted

EU Transparency Register: Redacted

From: "HOME INTERNET FORUM"<home-internet-forum@ec.europa.eu>

Date: Fri, Nov 8, 2024, 1:06 PM

Subject: RE: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024 - Ares(2024)7942846

To: "Redacted Redacted tiktok.com">

Cc: "Redacted twitter.com"<Redacted twitter.com>, "Redacted google.com"<Redacted google.com>, "Redacted fb.com"<Redacted fb.com>

Dear Redacted dear all,

Thank you for coordinating on this initiative already.

We are happy to extended the deadline until 15 January if this allows you to provide meaningful and targeted answers to the questionnaire that go beyond information available on your websites.

We are looking forward to your interesting contributions and engagement in this important workstream.

Kind regards,

The EUIF team



European Commission

Directorate-General Migration and Home Affairs

Directorate D: Internal Security

Unit D3: Prevention of Radicalisation

E: HOME-INTERNET-FORUM@ec.europa.eu

From: [Redacted] <[Redacted]@tiktok.com>

Sent: Thursday, November 7, 2024 5:50 PM

To: HOME INTERNET FORUM <HOME-INTERNET-FORUM@ec.europa.eu>

Cc: [Redacted]@google.com; [Redacted]@fb.com; [Redacted]@twitter.com

Subject: Re: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024

Dear EUIF team,

Thank you for looping me into this thread, and for your follow up email.

I am also copying some of the other companies, as we have just had a chance to catch up on this initiative. Would it be possible to get an extension on the feedback period, until January? We are currently juggling several pieces of work in the short lead up to the Christmas break, and will not be able to meet the 15th November deadline.

Kind Regards,

[Redacted]

--
[Redacted]

Redacted

Brussels

Tel: **Redacted**

Email: **Redacted**@tiktok.com

EU Transparency Register: **Redacted**



From: HOME-INTERNET-FORUM@ec.europa.eu

Date: Mon, Nov 4, 2024, 4:19 PM

Subject: [External] REMINDER - Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024

To: "HOME-INTERNET-FORUM@ec.europa.eu"<HOME-INTERNET-FORUM@ec.europa.eu>

Dear all,

This is a small reminder for the EUIF questionnaire on algorithmic amplification.

In case of questions please do not hesitate to reach out.

We are looking forward to your replies.

Kind regards,

The EUIF team



European Commission

Directorate-General Migration and Home Affairs

E: HOME-INTERNET-FORUM@ec.europa.eu

From: HOME INTERNET FORUM

Sent: Tuesday, October 15, 2024 9:38 AM

To: HOME INTERNET FORUM <HOME-INTERNET-FORUM@ec.europa.eu>

Subject: Informal Focus Group - Next Step - Questionnaire - DDL: 15 November 2024

Dear participants,

Thank you for your participation in the first meeting of the informal focus group on algorithmic amplification.

As an outcome of the meeting, the group agreed to create an overview of measures taken by companies since the EUIF study on algorithmic amplification (July 2022) to prevent the spread of borderline, violent extremist and terrorist content online and inform about existing commitments under the Code of Conduct on Hate Speech and Code of Practice on Disinformation. This overview could serve as set of best practices to inspire and inform.

If you are a company, we kindly ask you to fill in the attached questionnaire and send it back to us by **15 November 2024**. Please do not hesitate to reach out in case of questions.

Based on this overview and own research, ISD can identify gaps and propose potential improvements for algorithmic recommender systems to fight illegal and harmful content online.

We are looking forward to your replies.

Kind regards,

The EUIF team



European Commission

Directorate-General Migration and Home Affairs

Directorate D: Internal Security

Unit D3: Prevention of Radicalisation

E: HOME-INTERNET-FORUM@ec.europa.eu

Exhibit 37

DG HOME

EU INTERNET FORUM: STUDY ON THE ROLE AND
EFFECTS OF THE USE OF ALGORITHMIC
AMPLIFICATION TO SPREAD TERRORIST, VIOLENT
EXTREMIST AND BORDERLINE CONTENT



This is the Final Report of the study commissioned by DG HOME, drafted by Fincons, Trust Lab and Tremau, in the context of the EU Internet Forum.

Abstract

The European Commission (DG HOME) commissioned this study for the European Union Internet Forum. The study examines the degree to which the five leading social media platforms in the EU, and in particular their recommender systems which proactively present content to users, algorithmically amplify terrorist and violent extremist content to online users. The study also examines the extent to which recommender systems amplify Borderline content, such as some forms of hate speech leading to violent extremism, disinformation and other forms of legal, but harmful content, that could lead to the creation of online filter bubbles, and recruitment and radicalisation of online users. The study analyses Facebook, Instagram, TikTok, Twitter and YouTube, covering eight markets/languages in the EU. The study was conducted from July 2022 to May 2023. Prior to this Final Report, the team submitted an Inception Report and two Interim Reports that complement this report.

Executive Summary

The European Commission (DG HOME) commissioned this study, performed by Trust Lab in collaboration with Tremau and coordinated by Fincons Group, which examines the degree to which five leading social media platforms and their recommender systems¹ amplify harmful Terrorist and Violent Extremist (TVE) content and “Borderline” content² to online users.

The Report reflects the study of Facebook, Instagram, TikTok, Twitter and YouTube, that operate in eight markets/languages, namely, Arabic, English, French, German, Italian, Polish, Russian and Spanish. The study looked at: (1) the extent to which a platform’s recommender system algorithmically amplifies TVE and Borderline content to users, and the impact of the dissemination of such content on a user’s journey to radicalisation; and (2) the extent to which a platform’s recommender system leads to the creation of online “filter bubbles”³ and the impact on a user’s adoption of violent extremist beliefs.

This study was a collaboration between Trust Lab, Professor Theodoros Evgeniou, Professor of Decision Sciences and Technology Management at INSEAD, who researched the risk of radicalisation linked to the dissemination of TVE content, and Fincons Group which handled project coordination.

¹ Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content recommendations to online users. A key component of the recommender system is its recommender algorithm that determines the content a user will be served.

² As noted by the EU Internet Forum in its [Year in Review 2022](#): “Borderline content, also known as harmful, but legal content, is content that comes close to infringing on the community guidelines of social media platforms or laws regulating online illegal content. Some of the most common types of Borderline content identified in the EU are anti-establishment/anti-institutions, antisemitic, anti-trans, misogynistic, anti-migrants, racist, or against COVID-19 measures.”

³ “Filter bubble” refers to a homogeneous feed/recommended content, caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.

The European Commission's DG CONNECT has also been involved in the development of the study. The study was conducted from July 2022 to May 2023.

The study methodology is novel and discussed in Section 3.2 of the Report. Internal proprietary platform data (including algorithms), data, bots and Application Programming Interface (API) queries, which are common for this type of study, were not used. Trust Lab worked with human analysts to simulate motivated users intentionally seeking TVE content. The simulation was designed to ensure consistency and comparability across all platforms by the use of keyword lists, controlling for the amount of time exposed to each platform and repeating the experiments many times with different agents, thus increasing the likelihood of replicating the study's findings.

To address the six tasks defined by the study's sponsors, the following metrics were used:

- Findability: How easy is it to find harmful content, measured in the count of harmful posts during a one-hour interval?
- Removal Rate: What is the amount/percentage of harmful content removed by the platform?
- Removal Time: How much time did it take for harmful content to be removed by the platform?
- User Sentiment: What is the public (crowdsourced) perception of the appropriateness of harmful content, measured on a 5-point Likert scale?
- Amplification (filter bubble size): What is the amount of Bad Topic (TVE-related content, but not necessarily harmful content), Bad Content (TVE content), and Borderline content in a user's feed, measured as a percentage of the first 30 posts in the feed?

FINDINGS

Trust Lab's analysis found significant amounts of TVE content across all the major social media platforms as well as markets/languages. In addition, evidence of algorithmic amplification of TVE content on those platforms was also confirmed. The study identified significant differences in TVE-related performance and behaviour across all platforms and markets. The study's key findings include the following:

1. Evidence of amplification. The study validates the hypothesis that increased interaction with TVE content and Borderline content results in higher amplification of such content to users. All the platforms showed amplification of TVE-promoting content in their feeds. The percentage of TVE content in each platform's feed did not exceed 10% on average. Across all markets and platforms, the amplification of TVE content increased by 18% and Borderline content by 65%⁴. Twitter showed the highest level of amplification, YouTube ranked the second highest in amplifying TVE content, while TikTok showed the least. Regarding country and languages, platforms show the most TVE amplification occurring for Polish and German and for content related to Left-Wing political affiliation and younger Age Groups.

2. Findability Scores. Twitter had the highest findability scores of TVE content among the five platforms, and YouTube had the lowest findability score. Italian TVE content had the highest

⁴ This is the average percentage change between the first and third (final) amplification measurements across all languages and platforms. See the methodology, Section 3.2, for more details.

findability score of all languages when measured across all platforms. Violent Left-Wing TVE content had a significantly higher Findability score than other TVE types.

3. Removal Rates. More than 90% of the TVE content found by Trust Lab remained on the platforms by the end of the evaluation period 8 weeks later. TikTok removed more TVE content than the other platforms; Twitter and YouTube removed the least. Arabic TVE content was removed more frequently than the other languages, while Italian TVE content was removed less.

4. Borderline Content. Repeated cycles of user interaction and evaluation suggest that the amplification of Borderline content grows over time and at a higher rate than TVE content.

5. Variability across platforms and demographic blindspots. Each platform behaves differently in recommending TVE content with similar user features and behaviours. All platforms tended to recommend TVE at increasing rates as users interacted more with it, regardless of its TVE rating. Left-Wing politically oriented recommendations were far less likely to be removed by platforms than other groups. TikTok had a significantly low rate of success in amplifying TVE content to users, while Twitter and YouTube had the highest, with the former recommending most to Right-Wing and the latter recommending TVE content even to users with low rates of interaction. Ultimately, analysis based on user demographics and platform engagement data is still inconclusive to determine the full scope of how recommender algorithms operate on the back-end; more transparency and research is needed.

TABLE OF CONTENTS

1	OBJECTIVE OF THE FINAL REPORT.....	6
2	INTRODUCTION	6
3	APPROACH.....	7
3.1	THEORETICAL FRAMEWORK	7
3.1.1	Addressing Replicability	7
3.2	METHODOLOGY	11
3.2.1	Contributing Factors to Amplifying Harmfulness.....	12
3.2.2	Process.....	13
3.3	QUALITY.....	14
4	ANALYSIS	15
4.1	TASK 1 - MEASURING THE DISSEMINATION OF TVE CONTENT.....	15
4.1.1	Findability Main Highlights	15
4.1.2	Findability Findings	15
4.1.3	Removal Metrics Main Insights	16
4.1.4	Removal Metrics Findings	17
4.1.5	User Sentiment Main Insights.....	19
4.1.6	User Sentiment Findings.....	19
4.2	TASK 2 - MEASURING THE ROLE AND EFFECTS OF AUTOMATED DISSEMINATION OF TVE CONTENT..	21
4.2.1	Filter Bubble Main Insights	21
4.2.2	Filter Bubble Findings.....	21
4.2.3	Borderline Content	22
4.3	TASK 3 - ASSESS THE RISK POSED BY THE AUTOMATED DISSEMINATION	24
4.3.1	Risk Assessment Main Insights	24
4.3.2	Content Moderation Background	24
4.3.3	Risks Assessment.....	26
4.3.4	Risk Metrics	27
4.3.5	Mitigation Strategies.....	28
4.4	TASK 4 - COMPARE THE PHENOMENON BETWEEN PLATFORMS SHARING THE SAME BUSINESS MODEL	30
4.4.1	Comparison Between Platforms Main Insights.....	30
4.4.2	Platforms	30
4.4.3	Languages.....	31
4.4.4	TVE Type	31
4.4.5	Borderline Content	31
4.5	TASK 5 - ASSESSING THE RISK OF RADICALISATION DUE TO ALGORITHMIC AMPLIFICATION	33
4.5.1	Introduction	33
4.5.2	Analysis of Content Posts	34
4.5.3	Predictive Analyses of Sessions	38
4.5.4	Discussion.....	41
4.6	TASK 6 - PROVIDE GUIDANCE ON CONTENT MODERATION	43
4.6.1	Guidance on Content Moderation Main Insights.....	43
4.6.2	Systems.....	43
4.6.3	Tools	51
4.6.4	Third-party Services	52
4.6.5	Content Moderation Recommendations.....	53

5	CONCLUSIONS AND RECOMMENDATIONS	54
6	APPENDIX.....	56
6.1	PROJECT TEAM	56
6.2	KEYWORDS.....	58
6.3	EXTERNAL EXPERTS	62
6.4	TASK 1: FINDABILITY CHARTS.....	63
6.4.1	Findability per Platform	63
6.4.2	Findability per Language	64
6.4.3	Findability per TVE Type	65
6.4.4	Findability per Language / TVE Type	66
6.5	TASK 1: EXAMPLES OF ITALIAN VIOLENT LEFT-WING EXTREMIST CONTENT	67
6.6	TASK 1: EXAMPLES TWITTER CONTENT	69
6.6.1	Violent Right-Wing Extremism and Borderline Examples.....	69
6.6.2	Violent Left-Wing Extremism and Borderline Examples	72
6.6.3	International Extremism and Borderline Examples.....	74
6.7	TASK 1: USER ENGAGEMENT CHARTS AND P-VALUES	77
6.7.1	Average Number of Shares	77
6.7.2	Average Number of Likes.....	78
6.7.3	Average Number of Comments	78
6.7.4	Average Number of Followers	79
6.8	TASK 1: REMOVAL RATES FOR TVE CONTENT.....	80
6.8.1	Removal Rates per Platform.....	80
6.8.2	Removal Rates per Language	81
6.8.3	Removal Rates per TVE Type	82
6.8.4	Average number of Shares associated with Removed TVE.....	83
6.8.5	Average number of Shares associated with TVE that wasn't Removed	84
6.8.6	Average number of Shares associated with TVE that wasn't Removed broken down by language	85
6.8.7	Average number of Shares associated with Removed Borderline content.....	85
6.8.8	Average number of Shares associated with Borderline content that wasn't Removed	86
6.8.9	Average number of Shares associated with Borderline content that wasn't Removed broken down by language.....	87
6.9	TASK 1: REMOVAL TIME.....	88
6.9.1	Removal Time per Platform	88
6.9.2	Removal Time per Language	88
6.9.3	Removal Time per TVE Type	89
6.10	TASK 1: USER SENTIMENT METRICS.....	90
6.10.1	Severity Ratings by Platform.....	90
6.10.2	Severity Ratings by Language	91
6.10.3	Severity Ratings by TVE Type	92
6.10.4	Severity Ratings over Time per Platform	93
6.11	TASK 2: AMPLIFICATION CHARTS.....	94
6.11.1	Amplification Across all Platforms.....	94
6.11.2	Amplification Average Percent of Bad Content in Feed per Platform	95
6.11.3	Amplification Average Percent of Bad Content in Feed per Language	96
6.11.4	Amplification Average Percent of Bad Content in Feed per TVE Type.....	97
6.11.5	Amplification Percentage Change for Bad Content per Content Type.....	97
6.11.6	Amplification Percent Change for Bad Content per Platform.....	98
6.11.7	Amplification Percent Change for Bad Content per Language.....	99
6.11.8	Amplification per Platform.....	99
6.11.9	Amplification per Language	102
6.11.10	Amplification per TVE Type	106

6.11.11	<i>Amplification P-values</i>	108
6.11.12	<i>Interactivity and Amplification</i>	110
6.11.13	<i>Findability and Amplification</i>	110
6.11.14	<i>Engagement Ratio and Amplification</i>	111
6.11.15	<i>Removal Rate and Amplification</i>	112
6.12	BORDERLINE CHARTS	113
6.12.1	<i>Removal Rates per Platform</i>	113
6.12.2	<i>Removal Rate per Language</i>	114
6.12.3	<i>Removal Rate per TVE Type</i>	115
6.12.4	<i>Removal Time per Platform</i>	116
6.12.5	<i>Removal Time per Language</i>	116
6.12.6	<i>Removal Time per TVE Type</i>	117
6.13	TASK 3: LIST OF REFERENCES	118

1 Objective of the Final Report

The objective of this Final Report is to describe and document the work performed during this study and the outcomes of all completed Tasks. The report is extensively supported by examples from the measurement tasks, with case studies to showcase the role and effects of the algorithmic amplification and other visual aids. The majority of sections are supported by data visualisations in the form of charts, flowcharts, tables and other forms of data representation to present the results and tell the story in the most effective way.

2 Introduction

The existence of Terrorism and Violent Extremism (TVE) content online is one of the most concerning social media trends witnessed over recent years. This trend has been exacerbated by social media algorithms that provide curated experiences to users by recommending content that keeps users engaged with the platform for longer. The severity and amount of TVE is not contained to a few small corners of the internet. It is easily accessible on major online platforms and threatens personal health and safety, peaceful co-existence, and the well-being of all citizens⁵. Social media platforms are not always able to provide a safe online experience, and policies and enforcement practices can vary substantially across platforms.

This project aims to inform the European Union Internet Forum (EUIF)'s work related to the tools and measures used by online platforms to moderate TVE content. Furthermore, this project will provide insights and analysis into the algorithmic amplification of TVE content on five leading social media platforms (Facebook, Twitter, YouTube, TikTok and Instagram) across eight markets/languages (Arabic, English, French, German, Italian, Polish, Spanish and Russian). The overarching research question is how the five social media platforms compare in exposing motivated new users⁶ to TVE content. In order to answer this question, we designed a research method that allows for inter-

⁵ <https://cronfa.swan.ac.uk/Record/cronfa62902>

⁶ The term "New Users" refers to users who are intentionally looking for certain types of content.

platform comparisons and spans a broader range of platforms and languages than has typically been studied in the field of trust and safety research.

Our approach is to critically review the metrics collected relating to how social media platforms' machine learning-based content delivery systems amplify TVE content. The comparison across markets and platforms enables the ranking of individual platforms based on their performance, among other factors, to understand which platform provides the highest exposure and amplification of TVE content. Subject matter experts and industry veterans provide these recommendations to the European Commission based on assessments of the risks posed by the dissemination of TVE content and the potential for radicalisation through algorithmic amplification. Our expertise in trust and safety and online TVE and Borderline content shapes our guidance on content moderation.

3 Approach

3.1 Theoretical Framework

The goal of the study is to simulate the behaviour of motivated new users to research how social media platforms respond to users who perform targeted searches for TVE content. Motivated new users are new users on a given platform who have a targeted interest in TVE content (based on, for example, real-world interaction with TVE concepts) and who are looking specifically for that type of content on the platform.

The use of new accounts provides us with several advantages over using existing accounts. First, setting up accounts and pre-training them on 'benign' content would take a lot of time and resources that the current study scope doesn't allow for. Second, new accounts ensure that we have a more controlled research environment than we would have in the case of pre-existing accounts. In our current approach, we are aware of all the TVE-related content that agents find, and all the content that is recommended in the user feeds is captured and analysed.

If we pre-trained all the accounts on benign content, we would not only have to account for these differences in our analyses, but we would also have a harder time untangling the effects of the recommender system because of the different pre-exposure any of the accounts might have had during this pre-training phase. Therefore, starting with new accounts provides a clean baseline for comparison; even if the business models of the platforms under study are different, all our research on the platform is starting from the same entry point. In order to provide a quantitative analysis, such a common baseline is necessary.

Using new accounts also gives us the opportunity to test how quickly a recommender system will start recommending TVE content to these types of users. This information is relevant because the quicker the recommendation and amplification of TVE content occur, the more implications this has for mitigations on the platform's side. Thus, we are using the 'cold start' problem to our advantage here.

3.1.1 Addressing Replicability

Since user behaviour is at the heart of this study, the concern about the replication of findings must be addressed. In recent years, we have seen that social science research into human behaviour has

faced issues with the replication of findings⁷. This is because human behaviour is complex, and it is unwise and inaccurate to make generalisations from the results of a single study.

Behavioural studies like these will always suffer from issues with replication, at least to a degree. In order to minimise the chance of introducing unwanted variance while at the same time keeping the behaviour under study within the realm of realistic (albeit extreme/targeted), we used the following constraints:

- *Using a consistent understanding of TVE across platforms:* we have developed a set of policy definitions in accordance with the European Commission that covers a spectrum of TVE content and can be used cross-platform, enhancing the focus of our study. A consistent understanding is needed for this study, in part because platform-specific policies are inconsistent when compared to each other.
- These definitions are thus heavily informed by the existing platform definitions for TVE content but go beyond what the platforms currently define as TVE to more adequately capture the extent to which Borderline content exists on these platforms. This means that we expect content might not be removed from the platform if it is not covered by the platform's definition of TVE or that platforms might miss it during enforcement. This provides us with an opportunity to highlight the limitations of the existing platform definitions and provide mitigation strategies to address them.

Definition of TVE: Any media, including text, images, and videos, that promotes or glorifies terrorism or violent extremism or advocates for the use of violence to achieve political, ideological or religious goals.

- This category includes content that is the most explicitly promoting or supporting terrorism or violent extremism and which contains explicit calls to action or direct incitement to violence.
 - The content contains explicit calls to action or statements of support for terrorism or violent extremism.
 - The content contains inflammatory or provocative language that could be perceived as supportive of terrorism or violent extremism.
 - The content contains images or videos that are clearly promoting or supporting terrorism or violent extremism.
- ❖ **Violent Right-Wing Extremism (VRWE)** are acts of individuals or groups who use, incite, threaten with, legitimise or support violence and hatred to further their political or ideological goals, motivated by ideologies based on the rejection of democratic order and values as well as of fundamental rights, and centred on exclusionary nationalism, racism, xenophobia and/or related intolerance.⁸

⁷<https://arstechnica.com/science/2018/08/why-do-only-two-thirds-of-famous-social-science-results-replicate-its-complicated/>

⁸ The definition is provided by the DG HOME and incorporated into our methodology.

- ❖ **Violent Left-Wing Extremism** is a collective term for all efforts directed against the free democratic basic order that is based on treating the values of freedom and (social) equality as absolutes, especially as they are found in anarchist and communist ideas.
- ❖ Regulation Addressing Dissemination of **Terrorist Content Online** (EU 2021/784) (TCO) provides information for EU Member State competent authorities on reporting terrorist content. The full language of TCO can be found in the Official Journal of the European Union⁹. Under TCO, reasons for considering the material to be terrorist content are that such material:
 - 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.

Definition of Borderline content: This category includes 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)
- tactics used to manipulate users and amplify Borderline content leading to violent extremism, such as algorithmic amplification techniques that profit from biases in content sharing algorithms.¹⁰
- ❖ *Having only native speakers performing searches:* the underlying reasoning is that we want to understand language nuances in the platform's approaches to amplifying and/or moderating TVE and Borderline content and to get a better understanding of potential risk areas for specific languages.
- ❖ *Structuring the search tasks when it comes to keywords and search terms:* We used a fixed set of 60 keywords, translated for each language. Keywords were derived from existing research and our own evaluation and internal testing. The 60 keywords we chose are related to common terrorism themes across all markets.

This set is subdivided into three groups of 20 keywords for each type of TVE under study:

- Violent Right-Wing Extremism (Local/National and Ideological)
- Violent Left-Wing Extremism (Local/National and Ideological)
- International TVE (Religiously motivated terrorism)

⁹ Regulation (EU) 2021/784 of the European Parliament and of the European Council of 29 April 2021 on addressing the dissemination of terrorist content online, [2021], L 172/79.

¹⁰ This definition is taken from the EUIF Handbook on Borderline Content (2023).

Personas were assigned to one of three TVE types and only used their set of 20 keywords to provide a balance between being thorough and providing comprehensive coverage. Keywords were chosen according to Trust Lab's proprietary and representative selection criteria. Within this set of 20 keywords, there were a similar number of keywords for people, organisations, hashtags and slogans (or phrases). Agents were able to select keywords from within their set randomly but were not able to add new keywords. All keyword lists were translated by native speakers into terms that are culturally relevant in the local language¹¹. People's names were not translated, but hashtags and slogans may have been changed slightly in order to fit better with a specific language's nuances.

The keywords serve as a basis for the targeted searches. As such, this structure ensures that every search starts at the same place, with very few limitations with regard to how a user's journey subsequently unfolds after the initial search. Thus, we are recreating the 'rabbit hole' experience where a user might start with one video and end up exploring different videos.

The underlying reasoning for this approach is that having the same list of keywords for each language improves comparisons across platforms and languages. Some keywords will unearth more content than others in certain languages or on certain platforms, and being able to report on these differences will add a valuable component to our study. This would not be possible if we were to have different sets of keywords for each language or platform. Having different sets of keywords would also limit our ability to interpret any findings for metrics such as findability. This is because we would not be able to tell whether a certain keyword is a culprit or if a certain platform simply performs better in terms of concealing TVE-related content from its users.

¹¹ The list of English keywords can be found in the Appendix, Section 6.2.

3.2 Methodology

In this field of study, our methodology is comparatively unique. Instead of using bots or automated Application Programming Interface (API) queries, we use human agents (scouts) to simulate 'real' accounts. Their behaviour is designed to ensure consistency and comparability across platforms to enhance the likelihood of replication of the study's findings.

Human agent-driven investigations present certain limitations and challenges that have to be taken into consideration, such as human error, bias, and high costs associated with the method. Yet, we believe the agent-centric approach for data collection is best suited for the task at hand. One reason is that platforms' terms of service do not allow programmatic (automated) searching. Another reason is that the study aims to find out how user behaviour influences the content that is recommended to them (and, consequently, how that content influences the user), and this would not be possible with an automated approach. Lastly, and most importantly, it is the absence of primary data from the platforms themselves that place major constraints on the field of Trust & Safety research. Without access to platform internal data (access to the full corpus on content and behaviours) or a representative stratified sample (also needs to be provided by the platforms due to access restrictions) or a waiver for platform terms of service (TOS) that don't allow for programmatic access and searching, the results will be limited. In our discussions with the platforms, this point was raised multiple times, but broader data access was not provided. As a good sign for future research, the EU's Digital Services Act introduces provisions that allow access to data to researchers of key platforms and requires very large online platforms to disclose key data on the functioning of algorithms, a step very much needed to increase transparency and accountability for user safety.

The current approach is realistic yet constrained. In real life, people may not just use their accounts for searching for TVE content and might switch between different interests and topics, some of which are more innocent than others. A complete 'real life' approach would nevertheless introduce variance of a nature that would make it hard for our study to be replicable by others or to assess our results in a comprehensive manner. Therefore, we have chosen this approach while keeping these trade-offs in mind for future research.

Our study aims to understand how social media platforms' recommender systems algorithmically amplify TVE and Borderline content to users. Recommender systems, at their core, are algorithms that suggest relevant items to users - these could be products to buy, movies to watch, content to consume, and so on. Since there are many different recommender systems on different platforms, this could be an expansive endeavour. Some platforms recommend their users other accounts to follow/subscribe to, some platforms recommend content the user should consume next ('You watched this video, here's another one you might like'), and most platforms have separate 'Explore' pages that are exclusively filled with recommended content.

To ensure comparability across platforms and to keep the scope concise, we have chosen to limit this study to the recommender systems that populate the home feed of users. Feeds exist on all the platforms under study, while other recommender systems can be unique to one or several platforms and would, as such, not make as good a basis for comparison.

To compare platforms and markets directly, results need to be quantifiable, and metrics must be consistent across all platforms. To assess amplification, we have identified the following relevant metrics:

Table 3.2 - Metrics to assess amplification across platforms

Metric	Description
Findability	How many pieces of content a motivated user can find during a targeted search on the platform within a limited time window. This variable is a proxy for the amount of TVE content that is surfaceable on the platform.
Filter Bubble	A homogeneous feed/recommended content caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.
User Sentiment	Crowdsourced severity ratings of the content surfaced. These ratings will be on a Likert scale of 1-5 ¹² .
Removal Rate	How much content that is surfaced is being removed by the platform within a 8-week time period. This is passive removal; agents are not to report pieces of content as this would influence the algorithm.
Removal Time	How long it takes for a piece of content to be removed by the platform within an 8 weeks time period. This will be passive removal; agents are not to report pieces of content as this would influence the algorithm.

Findability is not the same as prevalence. We don't capture the total amount of content that agents come across, nor do we make any claims on how much TVE content actually exists on a platform. Findability is about how much content a person can find within a certain time frame.

Amplification in the context of our study can thus be seen as the increase in the amount of TVE content over time and the relationship between initial user searches and the amount of TVE content being recommended afterwards. Either one of these metrics indicates an amplification of TVE-related content.

In our study, we have been using two metrics ("Bad Content" and "Bad Topic") to measure the depth of filter bubbles. Bad Topic is content that is related to TVE topics but could be neutral, promoting or negative. This includes harmful TVE as well as EDSA¹³ content about TVE. Bad Content is synonymous with TVE content and what is also referred to as promoting TVE. It is also a subset of Bad Topic. We capture this distinction to evaluate whether the recommender systems recommend content to users based on the overall topic of TVE versus the actual content of the posts being consumed.

For the Moderation metrics (Removal Rate and Removal Time), Trust Lab used patented technology "Kaptix", which monitors each found piece of content for removals. Kaptix is an internally developed tool used to monitor whether social media posts are live on the platform.

3.2.1 Contributing Factors to Amplifying Harmfulness

In order to better understand which factors (behavioural or contextual) contribute to amplifying potentially harmful content, we have tracked both content metadata (number of likes, shares, and comments as well as the number of followers of the posting account) and behavioural data from our agents.

¹² A graphic depiction of the values on the Likert scale is available in Figure 4.1.5 of the report.

¹³ Educational, Documentary, Scientific, Artistic.

Agents were divided into High/Low interaction categories. Low Interaction agents only searched and watched content (for videos up to 5 minutes, regardless of video length), while High Interaction agents also liked and followed content. Agents did not comment or re-post/re-share the content. The influence of human interaction and contextual data on the harmfulness risk that recommender systems pose will be discussed in more detail in Task 5.

3.2.2 Process

Agents were assigned one or more personas. Within each language, 30 personas were created. Each persona had an account on each of the five platforms under study. This means that for each language, there were 10 Right-Wing TVE, 10 Left-Wing TVE and 10 International TVE agents, with 5 High Interaction and 5 Low Interaction agents per TVE Type. Agents were instructed to perform three Search Tasks and three Evaluation Tasks for each account. This allowed us to study the changes to the search results and feed recommendations over time.

All tasks were performed on mobile devices.

- **Search Task:** The Search Task started from a seeded keyword search. Agents were either in the Low or High interaction group and interacted with the content as per instructions. Agents were not allowed to add self-made keywords to searches. Agents were allowed also to browse their feeds to search for, interact with and capture TVE content according to the Interaction Type they were assigned to. Each Search Task was limited to one hour. We chose a one hour cutoff to ensure consistency among agents and due to resource constraints.
- **Evaluation Task:** The Evaluation Task started one day after the Search Task. This was to give the recommender systems enough time to update their recommendations. Agents were instructed not to interact with any content but to capture each post in their feed up to the first 30 posts.
- **Repetition:** Both Search and Evaluation Tasks happened 3 times to collect 3 data points for more accurate measurement and monitoring.

Table 3.2.2 - Number of personas per language

TVE-type \ Interaction type	Low Interaction	High Interaction
National / Violent Right-Wing Extremism (20 keywords)	5 personas	5 personas
International Terrorism (20 keywords)	5 personas	5 personas
Other / Violent Left-Wing Extremism (20 keywords)	5 personas	5 personas

With 8 languages and 30 personas per language, that means that we have 240 personas in total.

3.3 Quality

Results from the Search Tasks and Evaluation Tasks were subjected to a Quality Assurance Process to ensure that the content found and captured was indeed TVE content. For this work, we partnered with several international organisations that specialise in finding and analysing Trust and Safety-related content, as well as individuals from academia who study and work in the field of Terrorism and Violent Extremism¹⁴.

Table 3.3 - Quality Assurance

First Level Review	Search: Agent made an assessment of what is and is not TVE-related content. The agents have had policy training to make this distinction. Specialised QA agents reviewed 100% of the content. Evaluation: The agent performs a cursory labelling exercise on the web form. 100% of this sample is reviewed by specialised QA agents.
Second Level Review	A random sample of the complete dataset was used to analyse other metrics (not Findability) to control for the quality of the labelling performed by human raters. Due to the large amount of data that was collected and complexity of the labelling task, multiple rounds of labelling were necessary by qualified individuals on a smaller, cleaner sample. Based on the specifics of the data (amount of search stages, amount of platforms and languages under study), we arrived at a maximum sample of 7200. Since not all data combinations had 30 posts, the total sample size was 7074. The data was ensured to have the following: - Is the post related to TVE? (binary, based on average of rater's scale) - Is the post promoting TVE? (binary, based on average of rater's scale) - Is the post Borderline TVE content? (binary, based on average of rater's scale) - Content Appropriateness score (scale, crowdsourced) - Engagement Labelling (sourced from post data) To calculate the Findability scores, we used the complete set of content that was found during the Search phase. We relied on the first-level review to determine whether or not a piece of content was TVE.
Third Level Review	Content from the original dataset that had already been reviewed by Trust Lab experts and third-party academia experts was also included in the stratified sample as described above.

¹⁴ More details on the consulted experts can be found in the Appendix, Section 6.3.

4 Analysis

Due to the large volume of content we captured through our work we have chosen a stratified sampling method for the analysis of our Filter Bubble metrics and Removal Metrics. Findability scores were calculated using the complete set of data that was found during the Search Stages rather than a sample. User Sentiment data was also calculated based on the stratified sample.

4.1 Task 1 - Measuring the Dissemination of TVE Content

4.1.1 Findability Main Highlights

- **Twitter has the highest Findability score of all platforms analysed.** Since we don't know how specific characteristics of platforms (such as having an open API yes/no) influence moderation of content, we do not make assumptions on the potential impact of these characteristics on the amount of content we were able to find.
- **Arabic TVE has the lowest Findability score, while Italian has the highest.** Finding TVE content through search appears to be easier in Italian, while it's harder in Arabic. There are a lot of confounding factors that could contribute to this finding, so we cannot make assumptions about the cause.
- **Left-Wing TVE has a higher Findability score than other TVE Types.** This points towards a larger trend, where International and Right-Wing TVE content are more easily identifiable as extreme content, whereas Left-Wing TVE constitutes more variations across cultures and languages.

4.1.2 Findability Findings

Our analysis shows that Twitter has the highest Findability score (2.91) on the platform¹⁵, while YouTube has the lowest (1.71) among the platforms we analysed. This means that Facebook and Instagram, as well as TikTok, make up the middle on our scale. Findings for YouTube and Twitter are statistically significant. (See Figure A.6.4.1).

Moving to languages, Arabic TVE has the lowest Findability score (1.23), while TVE in Italian language has the highest (2.78). This means that finding TVE content through search is easier in the Italian language, while it's harder in Arabic. Both of these findings are statistically significant. (See Figure A.6.4.2).

For TVE Types, International TVE has the lowest Findability score (1.41), while Left-Wing TVE has the highest (3.01). That means that it's easier to find Left-Wing TVE, while it's harder to find International TVE by searching the platforms. These findings are statistically significant. (See Figure A.6.4.3).

Deep Dive: Violent Left-Wing Extremist Content

Since Violent Left-Wing Extremist content has a significantly higher Findability score than other TVE types, we wanted to explore this category of content in a bit more depth. The kind of content we find

¹⁵ Examples can be found in the Appendix B, Section 6.6.

when we look for Violent Left-Wing Extremist content is mostly supportive of anti-capitalism, pro-anarchism and pro-communism¹⁶.

These findings point towards a larger trend. Terrorist and Violent Right-Wing Extremist (TVRWE) topics, such as neo-Nazism, anti-Semitism, racism and white supremacy, frequently attract the attention of policymakers and journalists, which causes social media platforms to invest in technical and human resources to reduce the harm of such material. On the other hand, Violent Left-Wing Extremist content encompasses a range of topics that have entered modern political, social and cultural spheres without causing the same level of concern, such as anti-capitalism, general statements about the failure of modern-day institutions and the desire to learn about other forms of governance, such as communism and anarchism.

Deep Dive: User Engagement Metrics

In order to better understand how many people might have engaged with the TVE content that we surfaced, we performed an engagement analysis on four metrics: likes, comments, shares and the follower count of the accounts who posted the TVE content.

- TVE content on Instagram and YouTube is shared much less widely than content on other platforms. Content on Twitter was shared more widely. (See Figure A.6.7.1)
- TVE content on YouTube is liked more on average than other platforms. Facebook has the lowest number of likes for TVE content. (See Figure A.6.7.2)
- TVE content on YouTube is commented on more on average than other platforms. Comments on Instagram seem virtually non-existent. (See Figure A.6.7.3)
- TVE content on YouTube comes from accounts with more followers on average than on any other platform examined. (See Figure A.6.7.4)
- While YouTube has a low Findability score, the platform scores high on almost all engagement metrics, except for shares. This seems to indicate that the platform is in control of suppressing Bad Content from being found in search but that users have found other ways to access the content. (See Appendix, Section 6.7 for more details)
- Facebook and Instagram are generally at the bottom of the engagement charts. Especially Instagram scores low, having no shares and no comments on average for TVE content. (See Section A.6.7)
- Comparing the TVE vs non-TVE engagement averages shows that TVE content gets much lower engagement than non-TVE content, but these results are not statistically significant for most platforms. (See Appendix, Section A.6.7 for more details)
- It should be noted that the differences in engagement between platforms could also be due to the user interfaces of the platforms. It could be argued that YouTube and Instagram are less set up for sharing content than Twitter or TikTok, for instance. (See Appendix, Section 6.7 for more details).

4.1.3 Removal Metrics Main Insights

- **The platforms removed very few items** during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it, resulting in TVE being shared heavily on some platforms like Facebook.

¹⁶ Examples can be found in the Appendix, Section 6.6.2.

- **Twitter removes the least amount of TVE content.**
- **Arabic language TVE content is removed more often than other types of TVE content, while Italian TVE content is removed less.**
- **Terrorist and Violent Left-Wing Extremist content is also removed less than other types of TVE content.**
- Findability and Removal Rate are related to each other: the less likely it is that TVE content is removed, the more likely it is that we can find it. **It is, therefore, not surprising that the findings in this section resemble the findings from the Findability section.** Twitter removes the least amount of TVE content, and Arabic TVE content is actioned and removed more, while Italian TVE content is removed less, and Left-Wing TVE content is also removed less, than other types of TVE content.
- **Combining these insights seems to indicate that platforms tend to dedicate less effort to removing terrorist and Violent Left-Wing Extremist content and Violent Extremist and terrorist content in Italian language and instead focus more on removing violent Extremist and terrorist content in Arabic language, International and Right-Wing TVE content.** Since platforms still remove less than a third of the content marked as TVE Promoting, this seems to point towards inconsistent content moderation policies, in particular, less pronounced policies and enforcement focus towards markets and topics that garner less attention from the general public.
- **YouTube scores low on Removal Rate, which is an interesting reversal case:** The platform does comparatively well with suppressing TVE content in search without removing such content from the platform. See Section 4.2.2 for a deep dive into this.

4.1.4 Removal Metrics Findings

Our analysis shows that TikTok removed more TVE content than other platforms (11% of the content removed), while YouTube removed the least amount of content (3% of the content removed). Overall, platforms removed less than a third of the content marked as TVE promoting. These findings are statistically significant. (See Figure A.6.8.1).

Arabic language TVE content is more often removed than most other languages (10% of the content removed), while Italian language TVE content is less often removed than most other languages (3% of the content removed). These findings are statistically significant. (See Figure A.6.8.2).

TVLW content is less often removed than other TVE types: only 4% of the time, compared to 7% (Right-Wing TVE) and 8% (International TVE). This finding is statistically significant. (See Figure A.6.8.3).

The platforms removed very few items during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it. All Removal Time metrics are anecdotal because of the small number of removals overall. At each interval, the

percentage represents the amount of content that was removed out of the total removed within an 8-week period. (See Figure A.6.9.1).

Generally, German language TVE content was removed more quickly, while Italian TVE content was removed more slowly than other languages we've studied. (See Figure A.6.9.2) TVLW content also seemed to be removed the slowest, while International TVE content seemed to be removed more quickly. (See Figure A.6.9.3).

Deep Dive: Removal Metrics

The low rates of removal by themselves do not tell a very satisfactory story, so a deeper analysis was conducted to look into the impact that this is causing based on the data that was collected. The focus of this analysis was to contrast removals with user engagement and to review in particular, the number of shares. Shares were chosen because this is a much more deliberate action that also highly impacts the dissemination of TVE content than other forms of engagement¹⁷. It's worth noting that shares are most likely the least accurate for platforms like YouTube that have a significant presence on desktops because sharing can be just as easily done by copying the URL from the address bar. Also, sharing off-platform (for example, sharing a link on another social network) may not be trackable by the content owner.

Of the content that was eventually removed from platforms within the study period, TikTok TVE content was disseminated the most, with 120 shares per TVE post on average (see Figure A.6.8.4). However, of the TVE that wasn't removed by platforms Facebook has the highest number of shares per post 341 followed by TikTok with 283 on average (see Figure A.6.8.5).

The proportionally higher removal rates of TikTok contrasted with higher average number of shares suggests that TikTok TVE content is being consumed and spread at higher rates than other platforms. Thus, despite TikTok's higher performance on other metrics, it is struggling to keep up with its user base and high throughput even on high-harm content like TVE. This same trend appears in Borderline content except with higher shares counts (see Figure A.6.8.7).

Looking at how different languages impact TVE content that wasn't removed, French and Italian have more shares on Facebook while other languages are more prevalent on TikTok (see Figure A.6.8.6). Borderline content in German language, however, sees more shares on Facebook which is the only language that changes between TVE and Borderline content (see Figure A.6.8.9).

¹⁷ Ljungberg, J. et al, Like, Share and Follow: A Conceptualisation of Social Buttons on the Web, July 2017.

4.1.5 User Sentiment Main Insights

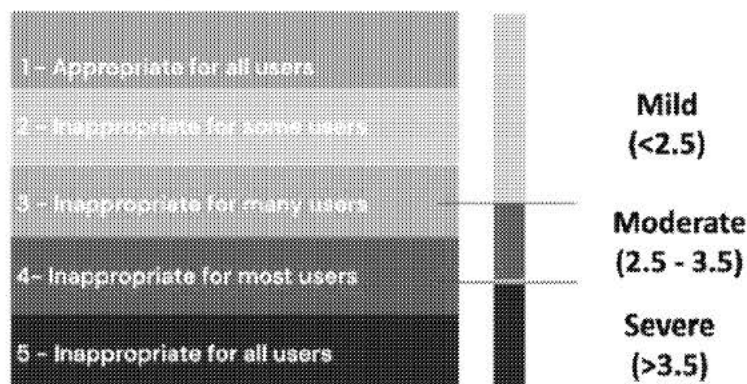


Figure 4.1.5 - Appropriateness score

- **Platforms are mostly similar** in terms of **severity of content based on user ratings**, indicating that severe content is spread roughly equally across the platforms. Italian content is rated as the most severe, while German is rated as the least severe. Most differences between languages and their severity ratings are statistically significant.
- **TVE content in Italian is rated as the most severe, continuing a trend that seems to indicate that Italian language TVE content is largely going unnoticed by platforms.** Other languages are also mostly statistically significantly different from each other, indicating that users of different languages have potential cultural differences when it comes to how harmful they judge TVE content to be.

4.1.6 User Sentiment Findings

Platforms are mostly similar in terms of how severe users rate their content, indicating that severe content is spread roughly equally across the platforms. (See Figure A.6.10.1).

Italian language TVE content is rated as the most severe (98% rated as severe), while German language TVE content is rated as least severe (43% rated as severe). Most differences between languages and their severity ratings are statistically significant. (See Figure A.6.10.2).

Right-Wing TVE content is viewed as less severe by users than other TVE types (78% rated as severe), which is a statistically significant finding. This is interesting and could be due to the fact that users are getting more used to seeing Right-Wing TVE content, but also to the fact that platforms are removing the more severe Right-Wing TVE content from their platforms more readily. (See Figure A.6.10.3).

Severity Over Time

Does searching for and interacting with TVE content influence the severity of the content that is subsequently found and recommended to users? A trend here could be an indication of amplification, not necessarily in the *amount* of content that a user sees, but in the *harmfulness risk* of the content that users engage with.

No clear trends emerge when looking at the severity ratings over time per platform. (See Figure A.6.10.4).

4.2 Task 2 - Measuring the Role and Effects of Automated Dissemination of TVE Content

4.2.1 Filter Bubble Main Insights

- **Taking all platforms, languages and TVE types together, there is an amplification of all content types over time.** This amplification is only significant for Bad Topic and Borderline content, but not for Bad Content.
- **All platforms show signs of amplification after the first Search phase, showing Bad Content in the user's home feeds, but platforms do not show any significant further amplification of Bad Content over time.** On average, no more than 10% of any feed consisted of Bad Content.
- **All languages show signs of amplification after the first Search phase, but only Italian shows significant further amplification of Bad Content over time.** French does show a significant further amplification of Borderline and Bad Topic content over time.
- **All TVE Types show signs of amplification after the first Search phase, but none show significant further amplification of Bad Content over time.**
- Overall, these findings seem to indicate that although there seems to be an amplification of Bad Content in the feed after showing initial interest by a user, this is not significantly further amplified when the user continues to engage with said content.

4.2.2 Filter Bubble Findings

Taking all platforms, languages and TVE Types together, there is an amplification of all content types over time. For Bad Content, the main focus of our study, our amplification metric increased from 4.4% to 5.2%, an increase of 17.5%. Looking at Bad Topic, which includes TVE and non-harmful content related to TVE topics, we saw an increase from 6.9% to 9.6%, which is statistically significant (+39.1%). Additionally, we looked at Borderline Content, which amplified from 3.3% to 5.4% (+65.1%), which is also statistically significant. (See Figures A.6.11.1).

Instagram is the only platform that shows a significant amplification across the three stages for Bad Topic content (+118%), as well as Borderline content (+150%). None of the platforms show a significant further amount of amplification for Bad Content, although Facebook doubled the amount of recommended Bad Content between the First and Third stage. TikTok showed no increase, and YouTube had a negative amplification (-23%); all not statistically significant. (See Figures A. 6.11.8a-e).

Italian and French are the only languages that show significant amplification of TVE related content over time. Italian is significant for Bad Content only (+365%). The French amplification is only significant for Bad Topic (+293%) and Borderline content (+685%), not for Bad Content. Polish has the highest average percentage of Bad Content in the feed (11%), Arabic has the lowest (2%). (See Figures A.6.11.9a-h).

None of the TVE types show significant amplification over time. (See Figures A.6.11.10a-c).

Higher interaction with TVE content does result in a higher initial percentage of Bad Content in the feed (8% for High Interaction, 1% for Low Interaction). (See Figure A.6.11.12).

Deep Dive: How Platforms Reduce TVE Content

When looking across the Findability and Filter Bubble metrics, we can compare how platforms choose to reduce the spread of TVE content. (see Figure A.6.11.13).

- **YouTube:** has the lowest Findability score but a high average percentage of Bad Content in the feed.
- **TikTok:** reflects the opposite trend, having a moderate Findability score and the lowest average percentage of Bad Content in the feed.
- **Twitter:** has the highest Findability score and also the highest average percentage of Bad Content in the feed.
- **Instagram:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.
- **Facebook:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.

YouTube seems to have tighter filtering of search results, while TikTok has tighter filtering of the recommendations in the feed. Relatively speaking, Twitter has the most on both dimensions. Facebook and Instagram appear to have moderate amounts of filtering in both search and feed.

These findings could be the result of YouTube's and Twitter's algorithms having stronger personalisation, combined with a lack of filtering. When platforms strongly personalise content based on past activity but lack adequate filtering of Bad Content, the filter bubble effect becomes most problematic. TikTok's algorithm is known to have strong personalisation effects¹⁸, but seems to have better filtering of TVE-related content than YouTube and Twitter.

4.2.3 Borderline Content

Special interest from the European Commission has been expressed with regard to the effect of the spread of certain types of Borderline content, such as disinformation and some forms of hate speech that can lead to radicalisation. This section will shine more light on how Borderline content plays into the analysis of the metrics that we calculated. It is content that does not meet the thresholds to be labelled as terrorist and violent extremist content, but that can still lead to violent extremism and radicalisation pathways.

Amplification of Borderline Content

- Overall, Borderline content has the lowest initial amplification of all TVE-related content types, but the percentage of Borderline content that is recommended to users has the highest amplification over time of all TVE-related content types.
- Instagram and Twitter are the only individual platforms that show a significant amplification in the amount of Borderline content that is recommended to users over time. Instagram shows 3 times more Borderline content when comparing the First to the Third phase.
- French is the only language that shows a significant amplification of the amount of Borderline content being recommended to users over time. 8 times more content is recommended to users in the Third phase compared to the First phase.
- None of the TVE Types shows a significant amount of amplification for Borderline content.

¹⁸ <https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity>

(See Figures A.6.11.8a-A.6.11.10c).

Removal of Borderline Content

- Overall, all platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any recommended content at all. Since we are comparing all platforms based on the same definition of Borderline content, this gives us a clear picture of which platforms are allowing more Borderline content on their feeds.
- Arabic Borderline content is removed more often during Search than most other platforms. None of the values for Evaluation are significant.
- Borderline content that could lead to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant.
- Removal Times for all TVE content and Borderline content are largely similar for all dimensions.

(See Section 6.12 for figures).

4.3 Task 3 - Assess the Risk posed by the Automated Dissemination

4.3.1 Risk Assessment Main Insights

- **Amplification: What is the prevalence of TVE content in the platform's feed?** Twitter was the platform with the highest level of amplification of Bad Content to the user, followed by YouTube, Instagram and Facebook. TikTok had the lowest level of TVE content amplification or harmfulness.
- **Risk Ratings.** We assigned a composite risk rating for each platform, based on an evaluation of three essential components: Findability, Amplification and User Sentiment. The risk ratings for the platforms mostly followed the percentage of TVE content found in the feed because the feed is where the bulk of the TVE content consumption likely occurs. The risk rating was consistent with removal and engagement rates, and there were fairly big differences among the platforms on this risk metric. Twitter had the highest risk rating of the platforms; TikTok had the lowest risk rating. Facebook, Instagram and YouTube had medium risk ratings.
- The average percentage of TVE content in the feed did not exceed 10%, even for motivated users, and did not seem to increase above 10%, even under repeated user interactions with TVE content. If platforms were only focusing on personalisation, this percentage would likely have been higher. This suggests that platforms either do not personalise much in general or that platforms limit personalisation, when TVE content, topic or Borderline content is detected. That said, when we compared the platforms on different dimensions, we observed clear differences among the platforms and further room for improvement, particularly for platforms such as Twitter with higher risk.

Recommender systems (RSs) are increasingly being used to support decision-making and improve user experiences across a wide range of industries and applications. However, it is important to acknowledge and effectively manage both technical and non-technical risks in order to reap their benefits and promote user safety fully. As requested by the European Commission, this report provides a comprehensive overview of recommender systems and the technical and non-technical risks associated with recommender system algorithms and outlines a range of mitigation strategies to help reduce risks and ensure best practices for the recommender systems and those who use them.

4.3.2 Content Moderation Background

Content moderation, or the process of monitoring, reviewing, and controlling user-generated content on online platforms to ensure compliance with community guidelines, terms of service, or content policies, is a fundamental part of the internet's infrastructure and success. Content moderation involves identifying and removing inappropriate, harmful, or offensive content, as well as addressing violations of rules and regulations set by the platform or applicable laws. The way a platform is designed and how users interact with it significantly influence the posting and interaction of user-generated content, consequently impacting how companies perform content moderation on their services (Grimmelman, 2015).

As discussed, platforms which rely on user-generated content often utilise recommender mechanics to deliver personalised content to users, deliberately surfacing tailored content most likely to keep the user engaged (O'Callaghan et al., 2014). Recommender systems have the potential to inadvertently recommend terrorist or extremist content to users who are not actively seeking such

material. This occurs because the algorithms prioritise user engagement without parallel automated objectionable content detection, allowing such terrorist or extremist content to be surfaced based on predicted engagement levels. Recent studies indicate that inadvertently exposing users to extremist content based on their expressed interest can lead to a reinforcement of extremist content in their feed; a consistent consumption of extremist content may contribute to a shift in their worldview (Edwards & Gribbon, 2013). Platforms face a challenging task in distinguishing between intentional bad actors and users who unintentionally engage in abusive behaviour, adding complexity to the moderation ecosystem. The following breakdown of prevalent filtering mechanisms and technological content moderation systems highlights the intricate balance between preserving the rights of users to freedom of expression and the challenging task of detecting and removing content in a scalable manner.

Content-Based Recommendation Systems

Content-based filtering is a technique used by online platforms, including user-generated content platforms like YouTube, to analyse a user's preferences and create a content engagement profile based on keywords and tags. This approach considers factors such as titles, descriptions, and tags to determine similarities and relevance when recommending videos, particularly for new or unseen content (Zhang, Lu & Jin, 2021). While content-based filtering offers scalability and independence from other users' data, it has limitations in the type of content it suggests to users. In the context of moderating terrorist content, platforms that rely heavily on content-based filtering pose challenges in effectively identifying and addressing such material. The focus on matching user preferences may overlook objectionable content that falls within the grey area of extremism. As the content becomes more extreme, the videos the system relies on for tags may begin to overlap between less extreme and more extreme content, making it less likely for a target audience to report such material to the platform. This increases the risks associated with content moderation and the potential for terrorist or extremist content to go undetected or unaddressed.

Although platforms basically work with some combination of recommender and flagging subsystems, these can be brought together in different ways, sometimes without a clear demarcation between the two. Also, while these subsystems draw upon the basic recommendation algorithms and flagging approaches outlined here to optimise for the metrics specified in section 4.4.2, there are many variations of these algorithms and metrics. Also, the practical pressures of running a business (e.g., generating revenue and retaining users, keeping infrastructure and machine costs low as well as addressing user and customer complaints and escalations) implies that these systems have accumulated numerous optimisations as well as business rules, making them exceedingly complex. As a real-world example, please refer to the recently outsourced Twitter recommender system.

Automated Flagging System

Automated detection systems utilise various techniques such as machine learning algorithms, near-neighbour algorithms, fuzzy hashing models, and MD-5 hashing models to identify harmful or platform-violating content, including both well-known and unknown terrorist and extremist material. These systems bring potentially illegal or policy-violating content to the attention of a team of content moderators, either directly employed by the platform or affiliated with a vendor company. Depending on the specific model employed, there are cases where reports are automatically closed if signals suggest the content is spam, does not violate the platform's Terms of Service, or has been previously reviewed by a content moderator. Automated detection systems play a crucial role in helping online platforms identify more severe content, initiating the cycle of moderation and policy enforcement on the platform.

User-Based Flagging System

User-based reporting systems rely on users to flag content that they believe violates an online platform's policies or is deemed objectionable. These reporting systems vary across platforms, often employing dedicated reporting forms or allowing users to directly flag content they find problematic while viewing it. User flags serve as a means for users to report content that may not have been detected by technical solutions, giving them a voice in the content moderation process. After users submit reports, the platform's team of content moderators reviews them and determines the appropriate actions to take regarding the reported content or offending user account. Many large online platforms deprioritise these user reports in favour of automated detection, but will no longer be allowed to do so under the new European Union's Digital Services Act, which places emphasis on users' right to report and appeal content they find objectionable, including terrorist and extremist content.

4.3.3 Risks Assessment

How can platform algorithms cause filter bubbles?

The term "filter bubble" is generally credited to internet activist Eli Pariser circa 2010. In Pariser's influential book, *The Filter Bubble* (2011), it was predicted that individualised personalisation by algorithmic filtering would lead to intellectual isolation and social fragmentation. To increase user engagement, social media companies may connect users with ideas they are already likely to agree with, thus creating echo chambers of users with very similar beliefs. The concern is that recommender systems may influence users to engage in progressively narrower content domains and in directions they might not otherwise have pursued. The EU Counter-Terrorism Coordinator has argued that the amplification of legal but harmful content may be conducive to radicalisation and violence because it normalises it and exacerbates polarisation in society.

The Global Partnership on Artificial Intelligence (GPAI) has also identified that recommender systems can amplify user bias related to TVE content, as follows:

The key issue for recommender systems . . . is that social media users are known to show small biases towards extreme content of various kinds, that act as another influence on the content items they engage with. For instance, they have a tendency to share political messages containing 'moral emotional expressions' (Brady et al., 2017; Brady and van Bavel, 2021), particularly negative ones (Crockett, 2017; Brady and van Bavel, 2021), messages that refer to a political 'out-group' (Rajthe et al., 2021), and messages that contain falsehoods (Vosoughi et al., 2018). If these biases persist while a user interacts with a recommender system, the system's repeated updates of its user model may lead the user towards messages containing increasing levels of negative political emotions, an increasing focus on political out-groups, and increasing amounts of misinformation – and potentially towards domains of violent extremism. Again, our earlier report (GPAI, 2021) presents these concerns and the studies that support them in detail.¹⁹

¹⁹ GPAI 2022, *Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions*. Report, November 2022. Global Partnership on AI.

Recently, the Global Internet Forum to Counter Terrorism commissioned a survey of the existing empirical literature on the issue of whether recommender systems promote extremist content and whether online users are “radicalising by algorithm”.²⁰ In the literature review, 10 of 15 studies demonstrated that recommender systems can promote extreme content.

Can bad actors exploit platform algorithms?

Terrorists could “game” flagging systems by experimenting and learning over time how to have their content bypass or circumvent either automated or manual flagging systems. According to certain researchers, the survival of certain TVE content on platforms relates to the way in which terrorist organisations have learned to modify their content to evade platform controls. Tactics include: breaking up text and using strange punctuation to evade platform tools that would search for keywords; blurring the terrorist organisation’s branding or adding the platform’s own video effects; mixing the terrorist organisation’s material with content from real news outlets; adding the branding of mainstream news outlets over the top of the terrorist organisation’s content; hijacking platform accounts; and posting tutorial videos to teach other terrorists how to do the same.²¹

Terrorists could also employ fake engagement and fake profiles to promote their content, and if undetected by platforms, recommender systems could unwittingly further promote such content. From gaming of views²² to crowdturfing and astroturfing²³, there are a variety of technical approaches to gain (perceived) popularity on social media, which are already known to be leveraged by foreign actors to influence operations. For instance, in one study, the BBC has reported that a network of Facebook groups was used in an attempt to change perceptions related to the ongoing war in Ukraine.²⁴

In this section, we presented many possible risks that can cause the promotion of TVE content on platforms. While we do not have access to internal platform documentation to understand how each individual platform’s algorithms and moderation systems work, we can assess risk based on externally observable metrics from this study. We do this in the next section.

4.3.4 Risk Metrics

The previous sections described the various risks that recommender systems present, but without access to each platform’s internal, proprietary algorithms and data, we were unable to quantify platform risk directly based on these factors. Instead, we evaluated the overall risk of each platform based on the data collected for this project, which could be observed outside-in. We considered three factors:

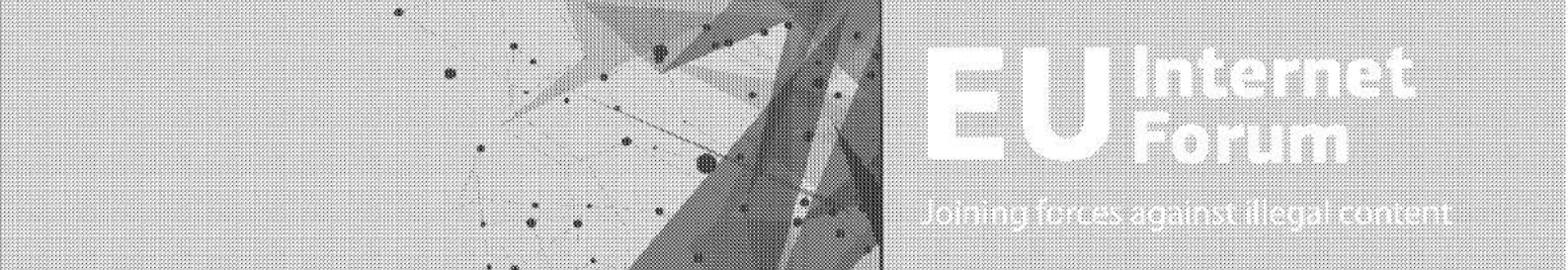
²⁰ Whittaker, J., Recommendation Systems and Extremism: What Do We Know? – Global Network on Extremism and Technology, Insights, 17 August 2022. <https://gent-research.org/category/insights/>.

²¹ Corera, G., *ISIS ‘still evading detection on Facebook’, report says*, BBC News, 13 July 2020. See also Nimmo, B. and Hutchins, E., *Phase-based Tactical Analysis of Online Operations (The Online Operations Kill Chain: A model to analyze, describe, compare, and disrupt threat activity from influence operations to cybercrime)*, Carnegie Endowment for International Peace, March 16, 2023.

²² The Flourishing Business of Fake YouTube Views. <https://www.nytimes.com/interactive/2018/08/11/technology/youtube-fake-view-sellers.html>

²³ <https://en.wikipedia.org/wiki/Astroturfing>

²⁴ Putin’s mysterious Facebook ‘superfans’ on a mission. <https://www.bbc.com/news/blogs-trending-61012398>



- User Sentiment: This is a proxy for the severity of TVE content found
- Findability: This is a proxy for the amount of TVE content that is surfaceable on the platforms
- Amplification: This is a proxy for the prevalence of TVE content in the feed.

We saw relatively minor differences in User Sentiment (see Figure A.6.10.1) among platforms. The platform with the highest level of Findability of TVE content (see Figure A.6.4.1) was Twitter, followed by Facebook, Instagram, and TikTok. The platform with the lowest level of Findability was YouTube.

Risk Outcome

We have assigned a composite risk rating for each platform based on our evaluation of three component metrics: User Sentiment, Findability, and Amplification. The risk ratings below incorporate information from the TVE content found in the feed and search and is consistent with removal and engagement rates.

Table 4.4.4 - Risk Rating per Platforms

Twitter	Youtube	Instagram	Facebook	TikTok
Highest	Medium	Medium	Medium	Lowest

The data we collected is based on simulating motivated users looking for TVE content and does not represent the experience of an average user of the platform. We also did not have access to internal proprietary platform data, which would have been needed for comprehensive measurement. That said, as previously discussed in the introductory section, it made sense to simulate and study the motivated user scenario. Also, the consistency observed for different metrics, such as feed prevalence, removal rates and engagement rates, suggests that the data is not an outlier.

The next section presents possible mitigation measures to consider, both based on observed data as well as conceptual risk factors discussed in previous sections.

4.3.5 Mitigation Strategies

There are a few different ways in which platforms can counter filter bubbles:

- By effectively detecting terrorist and extremist content, platforms can proactively filter such content from user feeds and search results. The **Global Internet Forum to Counter Terrorism (GIFCT)** is an Internet industry initiative to share proprietary information and technology for automated content moderation. As noted on the [GIFCT Wikipedia page](#), GIFCT maintains a database of perceptual hashes of terrorism-related videos and images that are submitted by its members and which other members can voluntarily use to block the same material on their platforms. The material indexed includes images, videos and will be expanded to include URLs and textual data such as manifestos and other documents.

- Filter bubbles and echo chambers can be countered by having recommender systems explicitly optimise for content or opinion diversity. There is evidence suggesting that personalisation algorithms can enable exposure to “ideologically cross-cutting content” from outside self-selected echo chambers. (Flaxman et al., 2016). It has also been suggested that they can also be leveraged in improving the automated detection of radicalisation and for facilitating targeted counter-messaging interventions to combat radicalisation. (Schmitt et al., 2018).²⁵ Recent literature has also suggested that algorithms can be the cure for algorithmic filter bubbles.²⁶
- Platforms should ensure that the algorithms themselves don’t have ideological bias. A recent Twitter study (“the most comprehensive audit of an algorithmic recommender system and its effects on political content”) indicated that the political right enjoys higher amplification compared to the political left. Our own data presented in previous sections also examines differences between Right- and Left-Wing TVE content both in terms of discoverability and moderation. Armed with such data, platforms can take a deeper look at their own algorithms and implement necessary mitigations.

²⁵ Wolfowicz, M., Weisburd, D. and Hasisi, B. (2021), Examining the Interactive Effects of the Filter Bubble and the Echo Chamber on Radicalization, *Journal of Experimental Criminology* (2023) 19:119-141 at 136. <https://doi.org/10.1007/s11292-021-09471-0>

²⁶ Gao, C. et al., Counterfactual Interactive Recommender System (CIRS): Bursting Filter Bubbles by Counterfactual Interactive Recommender System, *Computing Research Repository (CoRR)* in arXiv, abs/2204.01266 (2022).

4.4 Task 4 – Compare the Phenomenon Between Platforms Sharing the Same Business Model

4.4.1 Comparison Between Platforms Main Insights

- Among the platforms, Twitter surfaced and recommended the most TVE content to its users and performed below other platforms in terms of TVE content removal.
- Although YouTube's Findability score was the lowest among the platforms, its removal metrics suggested it has room to improve on removing TVE content that exists on the platform. YouTube generally scored higher than the other platforms on user engagement (e.g., likes, shares, comments, number of followers) with TVE content. This seems to indicate that YouTube was effective in suppressing TVE content from users in search but still enabled users to find TVE content through other means and interactions.
- TikTok performed moderately when it came to the Findability of TVE content, but of all the platforms, it recommended the least amount of TVE content to users. This suggests that TikTok may filter its feed more tightly than its search, resulting in no significant amplification of TVE content in the feed. TikTok was generally slow to remove TVE content; TikTok took more than four weeks to remove the content we identified as TVE. TVE content that the platform didn't remove was shared twice as much as TVE content that the platform did remove.

4.4.2 Platforms

Of the platforms, Twitter surfaced and recommended the most TVE content to its users. Twitter was lower than other platforms when it came to TVE content removal. We note that this study was conducted during a period when significant governance, policy and safety process changes were occurring at Twitter. As such, Twitter's results may be significantly different compared to earlier in the year.

YouTube has come under scrutiny during other research projects²⁷ and has since made improvements. Although YouTube's Findability score is the lowest among the assessed platforms, its Removal metrics combined with YouTube's engagement metrics for TVE content suggest the platform still has room to improve on removing TVE content that already exists on the platform.

In our study, YouTube recommended the most TVE content to its users; Twitter was second. It should be noted, however, that all the platforms had weak Removal rating scores, removing 10% or less of the content we marked as TVE-related.

YouTube generally scored higher than the other platforms on user engagement with TVE content. This seems to indicate that YouTube is effective in suppressing TVE content from users in search. But, it still enables users to find TVE content through other means and interactions.

²⁷<https://www.technologyreview.com/2020/01/29/276000/a-study-of-youtube-comments-shows-how-its-turning-people-onto-the-alt-right/>

<https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content>

4.4.3 Languages

Italian TVE content is an interesting case with regard to many of the metrics considered. The Findability score for Italian TVE content was one of the highest, and removal metric scores were in the lowest category. Under the study, Italian users in the sample were also the ones who rated their TVE content as the most severe of all the languages. This suggests that Italy may not be a high-priority market for platforms in terms of Italian-specific automated detection methods, Italian moderators, or other measures integral to a Trust and Safety enterprise.

TVE content in Arabic has long generated special attention from the platforms' moderation teams. Its Findability score is the lowest. Significantly, Arabic TVE content, especially International Arabic TVE content, is more often removed than content from other languages. Since the Paris attacks in 2015, there has been more focus from policymakers and media outlets on social media platforms to police the spread of International TVE from Arabic-speaking countries and regions, which could explain this finding.

4.4.4 TVE Type

Left-Wing TVE content treatment stands out. It's easier to find Left-Wing TVE content on the different platforms, especially in the Italian language. Left-Wing TVE content is also the least often removed. Apparently, this type of TVE content is also not as readily on the radar of platforms, possibly because Left-Wing TVE content covers a broad spectrum of content that is not TVE related per se.

This can be contrasted with Right-Wing and International TVE content, which receives more attention from policymakers and the media, thus incentivising social media platforms to create policies to moderate this type of content more readily.

Overall, one conclusion may be that the public attention to certain types of TVE content, often related to current events, plays a large role in what content gets moderated on platforms. If this is true, one could argue that this is an insufficient way of moderating TVE-related social media content. The platforms could certainly do more to be more consistent in their moderation practices (see also Task 3).

4.4.5 Borderline Content

When we look across the entire dataset, Borderline content appears to be amplified the most over time out of all the TVE-related content types that we've studied, even though Borderline content has the lowest amount of initial amplification.

Zooming in on different dimensions, such as platform, language, or TVE type, this trend is less pronounced. For example, Borderline content does not appear to be significantly amplified over time on any platform except for Instagram and Twitter. French is the only language where Borderline content was significantly amplified over time. Borderline content, potentially leading to terrorist and Violent Right-Wing Extremism, was not significantly amplified over time.

Platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any

recommended Borderline content. Borderline content in the Arabic language is removed more often during Search than other platforms, but none of the values for Evaluation are significant.

Borderline content potentially leading to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant. These findings are similar to other TVE-related content, which seems to show that the policies of the platforms do not place special focus on Borderline content.

4.5 Task 5 - Assessing the Risk of Radicalisation due to Algorithmic Amplification

Assess the level of risk of radicalisation to violent extremist ideologies and terrorism linked to the voluntary or involuntary automated dissemination of terrorist and violent extremist content through the use of machine learning-based algorithms or less sophisticated techniques.

4.5.1 Introduction

We employed various methods to analyse the data from the experiment and build models to predict whether a given social media platform user's demographics and behaviour could relate to TVE content amplification. In this section, we explain our process of analysis, and key results and briefly conclude with some final insights and recommendations.

Before proceeding, we note that machine learning processes are by nature exploratory (e.g., from selecting an appropriate way to organise the data, to setting up prediction problems and of course, to the selection of methods to use and the final analyses). Given the complexity of the data – and the problem – there was, naturally, significant exploration along the “machine learning pipeline” (from data structuring to final insights). We only discuss the processes, methods and results we found to be most appropriate.

While there are consistent observations across the different approaches discussed, each also provides some specific insights. Overall, consistently across all approaches outlined below, the results indicate that there is some predictability about whether TVE content is presented to users (during the Observation phase – see below).

As described in the previous sections, the experiment was carried out by having real agents impersonate generated user personas with different demographics (age, gender, country of origin and language), extreme political orientations (Violent Right-Wing Extremism, Violent Left-Wing Extremism, or apolitical/international extremism), and behaviours with TVE content (high interaction or low interaction). Over the course of three separate sessions (Search Stages 1, 2, and 3), users logged in and directly searched for and engaged with TVE content (the Interaction Phase) and, on a separate occasion, would log in again to observe the resulting content presented/recommended to them by a platform (the Observation Phase). The TVE content discovered in either phase was recorded (such as date, description, URL, etc.), its relation to TVE was rated by multiple experts (Borderline, promoting, or relating to TVE), and its popularity noted (account followers and the comments, likes, and shares for a given post). The raw data was organised into rows, with each row representing a particular post that was either found during the interaction or recommended during observation by a given persona on a given platform during a given session/login.

We study the effects that demographics, deliberate TVE searches, and the temporal aspect of doing so over multiple instances have on the platform algorithm's tendency to amplify such content. The process outlined below involves a preliminary exploratory analysis of the data using statistical descriptions and clustering by recorded content posts, followed by predictive modelling for TVE presence by session (login) for the Observation sessions – while using data also from the interaction sessions. Of course, if there is no (or weak) “signal” in the data, for example, for some platforms, one cannot make inferences reliably. Overall, the data did prove to be challenging to analyse.

Data Preparation

Data Cleaning

The data from the experiment's implementation contained very few instances of missing data. For example, only 54 (0.76%) out of 7074 rows (selected content posts across all sessions) used for the analysis contained discrepancies that rendered them unusable. This small fraction allowed us to directly remove those observations without hindering the quality of the analysis. Different features included in the data set were useful for different sections of this report, but overall the features kept for the analysis were the aforementioned demographic, behavioural and political data on the personas and the popularity and TVE ratings of the content they either found (Interaction phase) or were presented/recommended (Observation phase).

Encoding

Encoding is a process of converting data from one format to another. For example, it is used to convert categorical data (like gender or country) into numerical data so that it can be used in statistical models. There are many different types of encoding techniques, but two of the most common ones are ordinal encoding and one-hot encoding.

- *Ordinal encoding* is a method of encoding categorical data where each category is assigned a unique numerical value. This encoding is especially useful for including information about the relative significance or order of categories with respect to each other.
- *One-hot encoding* is a method of encoding categorical data where each category is represented by a binary column in the dataset. For example, if we use the Gender feature, 'male' could be labelled as {0,1} while 'female' as {1,0}. This technique is useful when there is no natural order or hierarchy among the categories, and each category is equally important.

The appropriate encoding method to use depends on the type of data. As we progressed through the analysis, we used different forms of encoding depending on the method implemented, which is mentioned in its relative section below.

4.5.2 Analysis of Content Posts

Statistical Description

We first performed exploratory analyses to determine the significance of each data feature collected and get an overview of each platform's frequency of TVE content dissemination. Once we have an overview description of the data, we can move to more complex analyses.

We start with statistics, also considering the temporal nature of the three cumulative search stages. We first map the occurrence of TVE-scoring content (Borderlines, Promotes, and Relates to TVE) based on the given Search Stage, Platform, and Interaction Level during the Observation Phase, meaning the content that was recommended to the persona after a period of engagement (*Figure 4.5.2a*).

This initial analysis indicates the following observations:

- There is overall (but different across platforms) a higher rate of TVE in the final third observation stage than the first two stages.
- There is a significant difference in the observed TVE content for Low-Interaction personas versus High-Interaction, but less so for YouTube.

TikTok appears to have overall very low TVE content, regardless of one's Interaction Level, with almost no instances of TVE content being present in the Observation Phase.

Finally, we also noted from the coefficients of all three TVE score categories that *Relates to TVE* was an appropriate category that encapsulates the average score of *Borderlines* and *Promotes TVE*. We, therefore, focus on that metric.

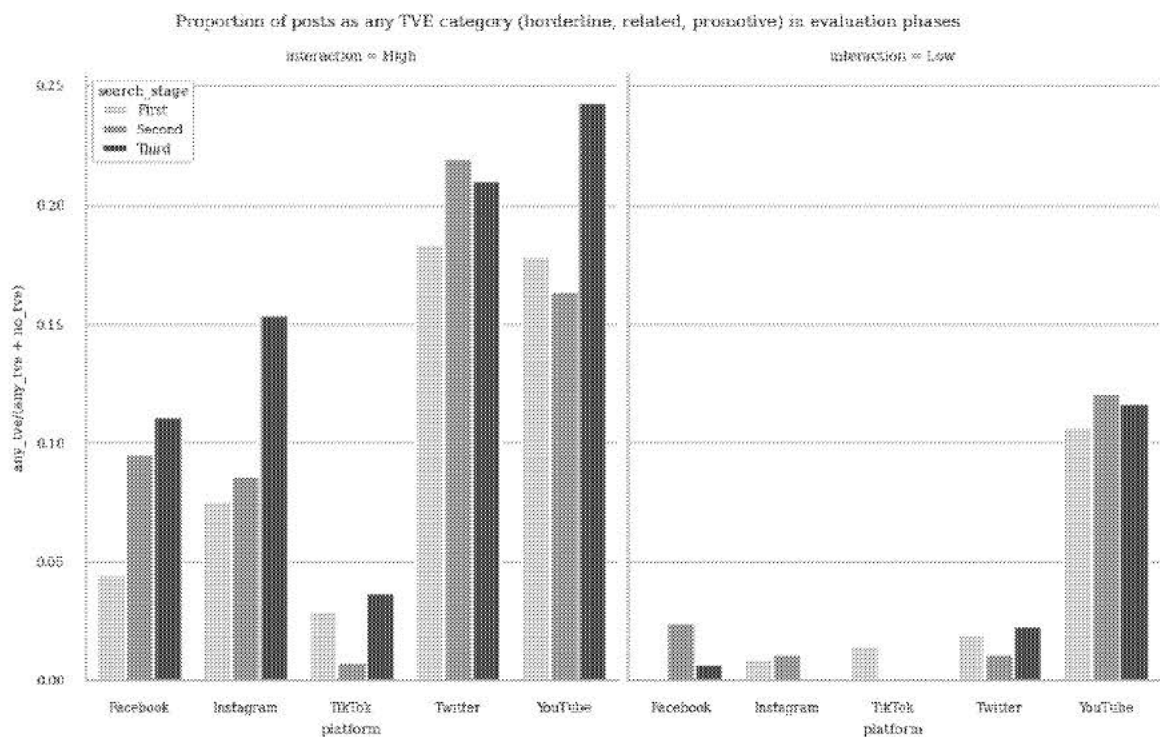


Figure 4.5.2a - Prevalence of all TVE categories, grouped by platform and divided between high and low levels of persona interaction

To further explore potential relationships, we plotted two versions of groupings of our significant demographic features and the TVE score. These features' importance is shown below (Figure 4.5.2b; Figure 4.5.2c).

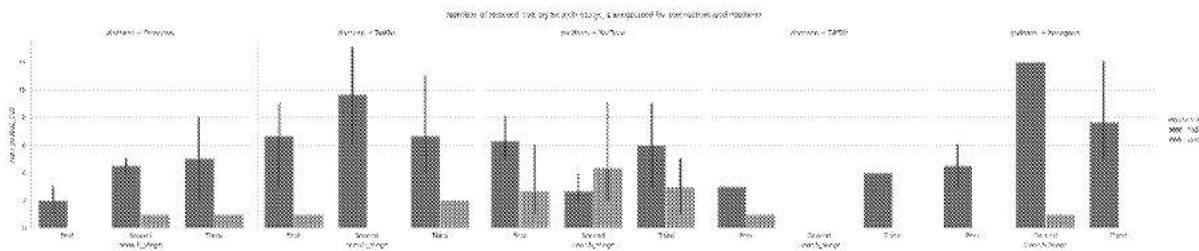


Figure 4.5.2b - Prevalence of related TVE across platforms, grouped by level of persona's interaction per search stage

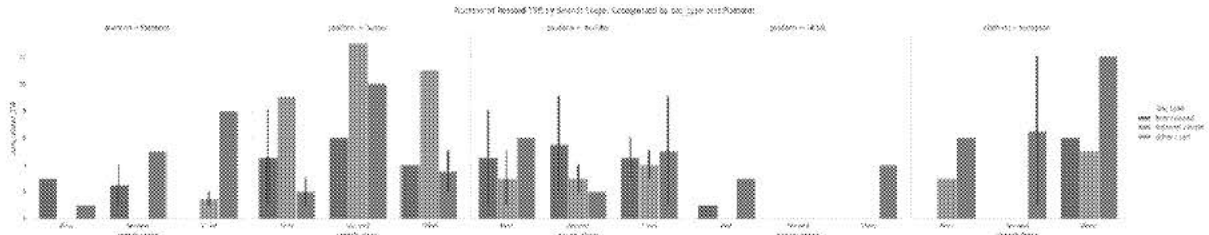


Figure 4.5.2c - Prevalence of related TVE across platforms, grouped by persona's political orientation per search stage

The results of these graphs display similar observations as the first statistical description (Figure 4.5.2a), with TVE presence generally increasing as the sessions accumulate. Twitter seems to cater most to recommending Right-Wing users TVE content while (again) TikTok has little to no instances of TVE content. Three observations can be made about the different platforms:

- YouTube's TVE content does not depend as much on one's interaction level, as above, or political orientation.
- As above, TikTok has overall little TVE, and while it displays relatively (across stages) more TVE content in the initial stage for High Interaction level personas and Left-Wing orientations, this is quickly hindered in subsequent phases. This could be indicative of a more robust internal system against TVE content amplification – but given the limited volume one cannot reliably make inferences about how the algorithms may work.
- Instagram and Facebook data are in agreement with the possibility that the platforms may have curtailing mechanisms against Right-Wing and International TVE content, but Violent Left-Wing Extremist users appear to “succeed” in gradually being presented with an increasing amount of TVE content over time.

We take this exploratory analysis further by implementing a machine learning methodology, clustering, to look for more patterns that may exist within the data of each platform based on the content posts recommended to a given persona.

Clustering by content posts

K-means clustering is a type of unsupervised (meaning the computer is not given labelled data and must find any hidden patterns on its own) machine learning algorithm that is used to group similar data points into ‘clusters’. The goal of k-means clustering is to partition a dataset into ‘k’ number of clusters. The algorithm works by randomly selecting points from the dataset as the initial cluster centres, then assigning all other data points to the nearest centroid based on its distance, and then iterating multiple times until convergence to clusters. We used ordinal encoding for this method and performed some feature engineering as described below to create the clusters of each individual social media platform by the content posts discovered during the Observation Phase of the study.

Feature engineering

As data collected is not always immediately obvious as to its relevance to an analysis’ goals, some features can be “engineered” out of the raw data. We employed feature engineering in four instances to make certain variables more relevant to the problem of finding commonalities between the personas’ demographics and the TVE-scored content recommended to them:

- 1) The ages of personas were categorised by age groups (Young Adults ≤ 31 , Middle Aged 32 - 51, and Older Adult > 51). The purpose here was to gauge if any age-related significance could offer generational insight into the issue.
- 2) The personas' political leanings were combined with their levels of interaction with TVE content as *risk for radicalisation* (for example, a high-interaction, Right-Wing persona would be "High Risk Right-Wing"). The rationale for this feature was to see if a platform's efforts were all-encompassing of TVE or biased towards a particular group(s).
- 3) We combined the four data points on a reported piece of content's level of user engagement as a popularity as the weighted sum of averages of likes (40%), comments (30%), shares (20%), and followers (10%) ($P = \sum \text{Likes} \cdot 0.4 + \sum \text{Comments} \cdot 0.3 + \sum \text{Shares} \cdot 0.2 + \sum \text{Followers} \cdot 0.1$). Recommender algorithms not only take the user's engagement but also amplify more popular content, and we weighted the variables based on the likelihood of each interaction for the average user (i.e. people are more likely to like a post they see on their feed than share it). As clustering methods are sensitive when many features (dimensions) are used, creating this composite "score" helped reduce the data dimensionality.
- 4) Finally, we made a feature for a content's *proximity to TVE* by summing together the scores of its Relates to, Promotes, and Borderlines TVE. These three scores are the averages between each expert's individual rating for the content. The purpose here was to analyse these scores separately, as well as whether they had more significance when combined.

For this analysis, we structured the data as follows: we divided the data by each of the five platforms and looked at that of the Observation Phase and using eight features with ~715 rows per platform, each representing a posted piece of content that was recorded. The features included were the four engineered features (Age Group, Risk for Radicalisation, Popularity, and Proximity to TVE) as well as the persona's language, country, gender, and search stage. As clustering is known to be harder when the number of features is large, we limited the data only to the features above. We used the WCSS method outlined below to figure out the number of clusters naturally occurring within the data and then proceeded to use the "kmeans" R software package to produce the k-means clustering models for each individual social media platform.

We used the "within-cluster sum of squares" (WCSS) of the data as a method to determine how many natural clusters exist and found that number to be 4 for all five social media platforms. This means that the data points in each cluster are more like each other than they are to the data points in other clusters.

The centres, or means, for each data feature, give us insights into what exactly is the commonality between the data of a cluster. Each platform had one cluster of data points with the greatest frequency of high-scoring TVE content. It is also important to note that the clusters themselves had a very high rate of overlap between them, which indicates that while we can take away some insights, clustering is not comprehensive enough to make any substantial conclusions. That said, while all five platforms performed similarly, there are some noticeable variables that stood out between the clusters:

- 1) All platforms had (as in the previous section) more TVE content starting after the second round of the Observation Phase. The users affected most tended to be in the Middle Adult age group (32 to 50 years).

- 2) YouTube was an outlier platform whose four clusters all leaned toward higher TVE-scoring content regardless of the other attributes of the clusters. Clustering of posts may not be an appropriate method for this platform/data.
- 3) For the other platforms, the average Risk of Radicalisation for a persona was higher for those who had a High Level of interaction with TVE content and a Left-Wing political orientation, with a skew toward the Low-Interaction / International Risk category.
- 4) Twitter had the highest rate of TVE content for High-Interaction / Right-Wing users.
- 5) TikTok had (again) the lowest, almost no, instances of TVE content.

Finally, TVE content was – perhaps not surprisingly – *unpopular*.

Note that the results are consistent with the previous section. Some additional insights from this analysis are that Search Stage, Political Identity, Interaction Level, and Platform appeared to be relevant for TVE presence.

4.5.3 Predictive Analyses of Sessions

To see if the demographics (including political and interactive behaviours) of a persona are predictive features of the likelihood their “efforts” (during the Interaction phases) to view TVE content on a platform will be “rewarded” with more TVE recommendations is to look at these features as independent variables (the inputs, or ‘X’) and the TVE scores of content discovered during the Observation Phase as dependent variables (the outputs, or ‘Y’). This means that the problem can be formulated as one of prediction or analysing the relationship between a dependent Y variable and 1 or more X variables.

For this phase of analysis, we ran different predictive machine learning methods analysing the different *sessions* (not individual posts) of the experiment. This is important because recommender algorithms learn by taking what the user engages with every time they use the platform to curate more accurate recommendations that fit the user’s interests. As the personas search for TVE content on three different occasions (Interaction Phases) and intermittently log in to view recommended content (Observation Phases), our models below will try to predict the effect of persona demographics and cumulative login sessions on the TVE scores of recommended content in the Observation sessions.

For accurate prediction, it is necessary to do processing of the data to weigh the features or determine which features are the most significant in affecting the outcome (the TVE score). We employ various methods to do this, with the goal of being able to use these weighted features to accurately predict future instances of TVE recommendation on a platform with machine learning algorithms. We completed two types of prediction problems: regression—where we use features from the Interaction Phase sessions to predict the amount of TVE content in the Observation Phase sessions—and classification—where instead we predict whether an Observation Phase stage session has TVE content or not.

Predicting the Amount of TVE Content

The five regression methods used—linear regression, support vector regression, random forest, decision tree and gradient boosting—are supervised learning models, meaning that they use both the input as well as see the output (in our case, the TVE scores) data to formulate predictive algorithms then. One-hot encoding was used for these methods. Although each method has its unique strengths

and weaknesses, they all share the goal of making accurate predictions and modelling relationships between variables. In particular:

- *Linear and support vector machine (SVM) regression* are two approaches to fitting a line that best describes the relationship between predictive and target variables. For example, we can use linear regression to find the line that best describes the relationship between a user's age and gender and the TVE-related content they are recommended. With that line, we can make future predictions of a TVE score based on a given age and gender.
- *Decision tree* is a way to make predictions by creating a tree-like model of decisions and their possible outcomes. At each separation of the tree, a decision is made based on a unique feature, and each branch represents a possible outcome based on the feature chosen. In this experiment, a decision tree is also useful because it can handle our categorical demographic features.
- *Random forest* is a way to make predictions by creating many decision trees and then combining their predictions. Each decision tree is created by randomly selecting a subset of the features and a subset of the data points.
- *Gradient boosting* is a machine learning technique that combines multiple decision trees to make accurate predictions about future events. It works by iteratively adding new trees to the model, with each new tree focusing on the examples that the previous trees got wrong. This approach allows the algorithm to learn from its mistakes and improve its accuracy across training iterations.

We first tested these methods by trying out two different groupings of the variables; one group testing all demographic X variables, like for the k-means clustering, and the other using only the features that were the most significant from the clustering.

After encoding the data, we fit it to a given model mentioned above and test for its predictive accuracy by looking at the R-squared value. R-squared is a measure of how well a regression model fits the data. Higher values indicate a better fit. It is calculated as the ratio of the explained variance to the total variance (which is the variability of a set of data points around its average value). It is useful for comparing different regression models.

We grouped the X variables only by four features (search stage, political leaning, interaction level and platform), with the Y variable being related to the TVE score, which was representative of the averages of the other two types of TVE scores. Each method was performed on each platform. The resulting R-squared values are outlined below (*Table 4.5.3a*).

Table 4.5.3a - R-squared values of regression methods

Method	R-squared Score
<i>Linear regression</i>	0.317
<i>SV regression</i>	-0.101
<i>Decision tree</i>	0.010
<i>Random forest</i>	0.245

We further explored which features and values had a positive or negative impact on the TVE score (Figure 4.5.3a). The results confirm earlier analyses that suggested High Interaction and latter observation stages played the largest role in high-scoring TVE predictability (Figure 4.5.3b). The negative coefficients for 'platform Facebook' and 'platform TikTok' suggest that these platforms are associated with a decrease in the value of 'num_related_TVE', which means, on average, TikTok and Facebook platforms have fewer related TVE content compared to the other platform types.

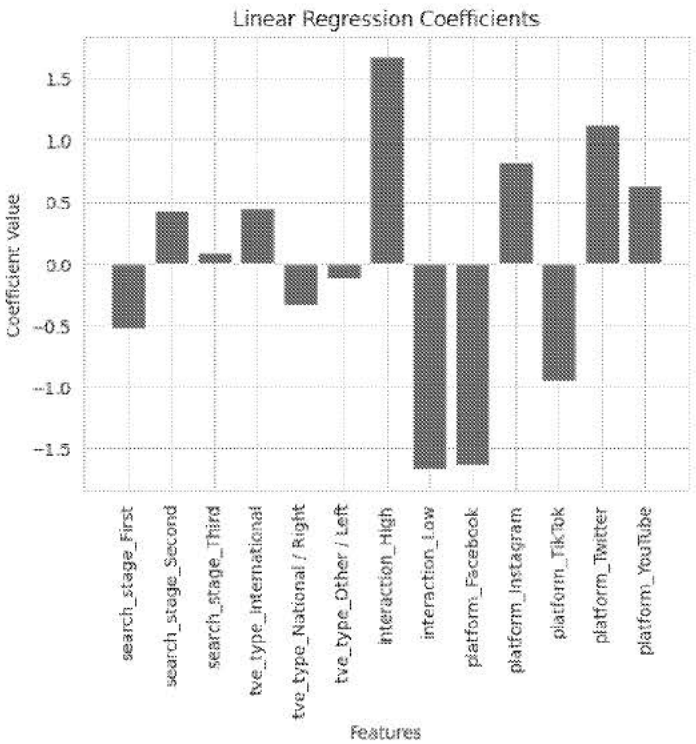


Figure 4.5.3b - Weights of different features used for linear regression

Both linear regression and random forest performed the best with R-squared values that hint that there may be some degree of a relationship between the four input variables and predicting TVE score of recommended content across the five platforms. However, predicting the TVE score of a login session during the Observation phase proved challenging – as the relatively low R-squared values also indicate. We therefore focused our analyses more on classifying whether an Observation Phase session/login had any TVE, hence on formulating and studying the relevant classification prediction problem, as discussed next.

Predicting the Presence of TVE Content

As for the regression analysis above, we used a sample of 1534 observation phase sessions, where a session is an instance of a persona browsing on a particular platform during one of the three interaction or observation stages. We only kept interaction sessions that were eventually followed by an observation session – and of course, observation sessions that had some interaction sessions before. As mentioned above, we assess each platform separately since we expect their algorithms to behave differently. We skipped analysis on TikTok since, as noted above, we found that this platform does not have a measurable increase in suggested TVE content in response to user searches (Figure 4.5.2a).

For the classification data sample, most of the data points correspond to sessions without TVE content (only 253 sessions out of 1534 have some kind of TVE-scoring content). Since a binary classification algorithm can achieve high accuracy by just making a constant prediction on this data, we needed to address the unbalanced distribution of TVE vs non-TVE sessions. We balanced the data by randomly up-sampling (creating synthetic copies) of the TVE content.

We trained several classifiers to predict whether an observation phase session contained any TVE content based on information from preceding interaction phase sessions that are related to the target observation session. We used a combination of categorical (e.g., persona country, political affiliation, gender) and numeric features (persona age, number of related, promoting, or Borderline TVE posts) from the earlier interaction phase sessions for the prediction of TVE presence in the subsequent observation phase session. We randomly split the data for each platform into a training (80%) and test (20%) set and applied ordinal encoding to the categorical features. This procedure is repeated 10 times; here, we report the results for a typical round. We implemented the classification methods using the Scikit-learn python package for machine learning and found that gradient-boosting tree-based methods had the best performance in all cases.

Table 4.5.3b – Prediction accuracy (% correct classified – 1 being 100%) for classification

Method	R-squared Score			
	YouTube	Twitter	Instagram	Facebook
<i>SV classification</i>	0.554	0.548	0.548	0.547
<i>Decision tree</i>	0.795	0.935	0.935	0.922
<i>Random forest</i>	0.807	0.925	0.925	0.922
<i>Gradient boosting</i>	0.831	0.914	0.914	0.898

Using the gradient boosting method – which was, as noted above, the most accurate – we also determined for each of the platforms which data features were the most important for prediction of TVE presence in the observation phase sessions (*Figure 4.5.3a*). The results indicate that largely consistent with the previous sections:

- The relevance of each feature regarding TVE content presence predictability varies by platform;
- High interaction in the Interaction Phase was predictive of the presence of TVE in the Observation phase for all platforms except for YouTube.
- Age was an important factor for all four platforms.
- Country, language and gender were also among the predictive factors, but as noted above, not similarly across platforms.

4.5.4 Discussion

First, we note that – perhaps not surprisingly – most of the content that was rated to be related to, promoting was almost entirely unpopular (meaning few likes, comments, shares, and followers). This may be simply because few people follow such content, or potentially the result of (recommender) algorithms effectively “supressing” this content – either scenario is possible. Given this, the popularity of content was not considered further in the final analyses discussed above. However, platforms do

not seem to fully suppress the exposure of users to such content when or after they come onto the platform, wilfully seeking it out.

Second, there are several consistent across-methods insights. First, TVE overall increased after the second Observation Phase, which may be due to algorithms. Second, there are differences across platforms. For example, TikTok had very low TVE content overall. Third, the presence of TVE in the Observation stages was predictable based on several features, both persona characteristics and – perhaps most important – the level of interactivity during the Interaction stages.

Finally, moving on to potential recommendations for social media companies, the results indicate that it may be important for platforms to look for the features that are most driving the probability for their users to be exposed to TVE content. On Instagram and YouTube, for example, those factors may be Age Group and Risk of Radicalisation. Companies can consider these factors as a starting point to address the issue of TVE content online.

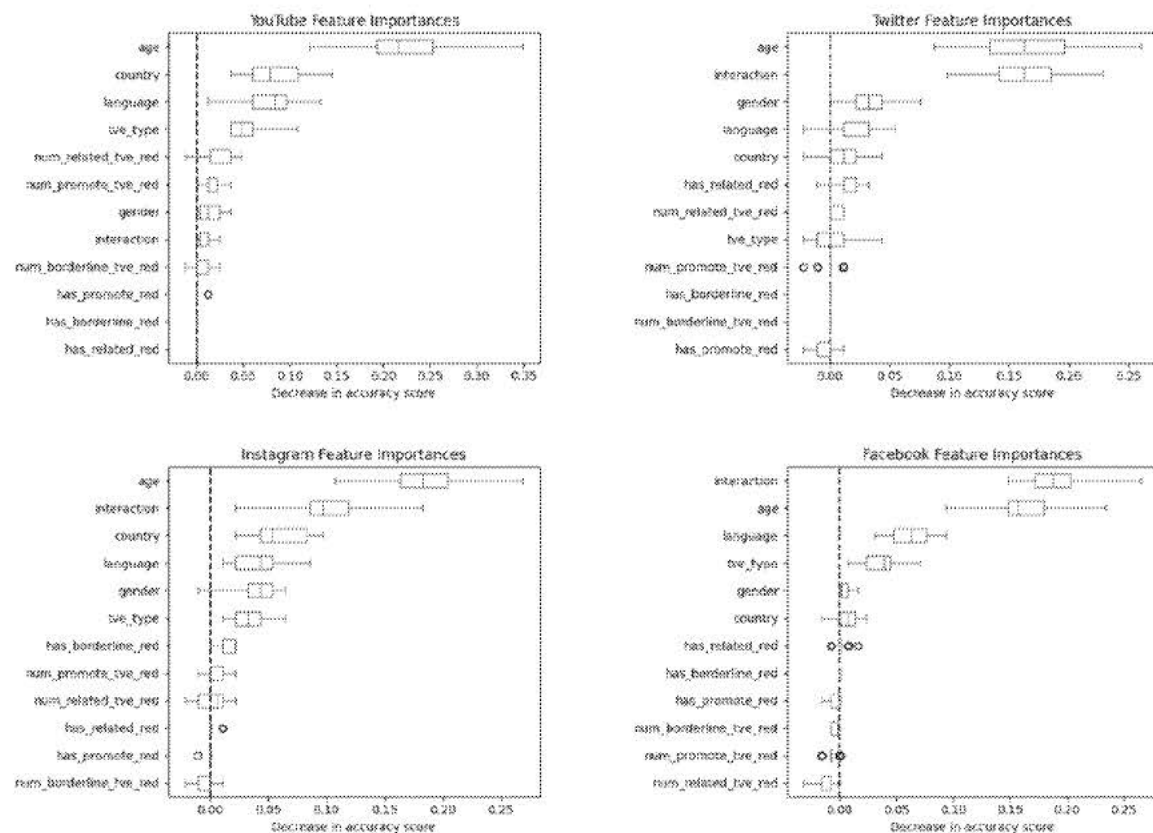


Figure 4.5.4a

These figures show the results of feature importance analyses on each platform to see which variables had the greatest impact on the accuracy of the models that predict the presence of TVE in the Observation Phase sessions.

4.6 Task 6 – Provide Guidance on Content Moderation

4.6.1 Guidance on Content Moderation Main Insights

The report outlines two different types of models that influence online users and the content with which they interact: content flagging models and content recommender models. Based on the study, it appeared that the content flagging models used by platforms were lacking compared to the content recommender models. Even if the recommender systems were reasonably adequate at preventing the most severe TVE content from being shown to users in their feeds, the question remains as to why the TVE content is surfaceable on the platform to begin with. Significantly, the platforms removed less than 10% of the content we marked as TVE-related, which indicates that the platforms need to improve their content flagging systems and policies.

4.6.2 Systems

TVE content moderation is unique to Trust & Safety efforts and platform specific. Companies operating online platforms have developed a variety of systems that aid in efforts to detect, remove, and punish content that violates a platform’s Terms of Service and Community Guidelines. The systems that platforms rely on can be roughly distributed on two axes, manual-automated and proactive-reactive. Coupled together, there are four system-based approaches to content moderation: Proactive Manual, Reactive Manual, Proactive Automated, and Reactive Automated.

Manual Systems

Manual systems require *human* intervention to *identify and review* potentially troubling material. Manual systems may rely on users to identify and report content, internal teams of employees who develop keywords or simple heuristics to identify and prioritise what content should be reviewed, and outsourced content moderators to review content.

Manual systems can be proactive or reactive in design. The key factor that distinguishes proactive manual systems from reactive manual systems is the extent to which review processes rely on user reporting (also known as “flagging” or “flags”). Proactive manual systems are not dependent upon users to find content for review. Instead, proactive manual systems rely on keyword searches of terms curated by company staff to find content related to targeted issues on a platform, for example, the use of a particular racial slur.

Table 4.6.2a – Proactive vs Reactive Matrix

	Reliant upon User Flagging?
Proactive Manual Systems	No
Reactive Manual Systems	Yes

As mentioned earlier, while not all manual systems prioritise the need for user flagging, they all require some degree of human intervention in the identification and review steps of Trust & Safety work.

Table 4.6.2b – Proactive vs Reactive Manual Systems

	Company Employees	Content Moderators (Raters)	Users (primarily through flagging)
Proactive Manual Systems	Design keyword-based lists for scrubs and sweeps	Review content identified from keyword-based lists to enact a range of options (e.g., removal, age-restriction, suppressing discoverability, issuing penalties)	Not a primary component of proactive manual systems
Reactive Manual Systems	Establish simple rules, or heuristics, to send flagged content to moderators	Review content surfaced from employees' rules to sort through user flags	Critical component of reactive manual systems

Proactive Manual Systems

For Proactive Manual Systems, company employees and contractors do not wait for users to find and flag content for review. Instead, ostensibly violative content can be surfaced through proactive manual searches (commonly referred to as “scrubs” or “sweeps”), which can be challenging, time-intensive, and ineffective in their reach. For example, proactive manual systems depend most frequently on the use of “keyword” matches where employees (or external entities, such as 3rd-party services entrusted by company staff) create and repeatedly modify lists of terms that are most troubling for a platform's staff.

The keywords used at any given time are shaped by a multitude of factors, including, but not limited to, the virality of content, proactive risk mitigation efforts (e.g., leading up to the anniversary of a known violent attack), and confirmed words, slogans, or phrases that are employed by particular TVE actors. In the event a staff member is concerned about, say, the anniversary of a violent attack, the employee may proactively pull content using keywords related to this event. From there, staff can review the material in question or send it to content moderators contracted by many platforms to conduct the majority of the moderation work. These keyword lists can be permanent (e.g., racial slurs, names of sanctioned individuals), curated for particular policy areas, or modified at time intervals most desirable for a platform's employees to add or remove terms. Keyword lists can also be used to automatically remove content that has a match with a term.

Keywords are not the only method of proactive manual systems, although they are quite popular due to the low level of sophistication necessary, high speed of deployment, and explainability. Less commonly, platforms may also choose to proactively place *any* content uploaded or created by its users under review or may even choose to place all newly created accounts under review as well. Proactive review, in this instance, aims to empower company staff and their outside moderation teams to more aggressively filter problematic, illegal, or undesirable material before users can discover and consume such content. It is important to note that there is no general monitoring obligation dictating the use of such systems.

Reactive Manual Systems

A reactive manual system is the traditional model of content moderation which leverages the wisdom and scope of crowds to surface bad content: user-reported content is sent for review by a company's staff or its outside moderators. There are several types of user flagging utilised in reactive manual systems.

In-Product Flagging

In-product flagging is almost universally applied across social media platforms. This feature enables users to submit complaints of content or users for review or incorporation into rules-based systems. (For example, a rule could be set that anything flagged by a user for being TVE content can be given higher priority by content moderators or automatically removed.)

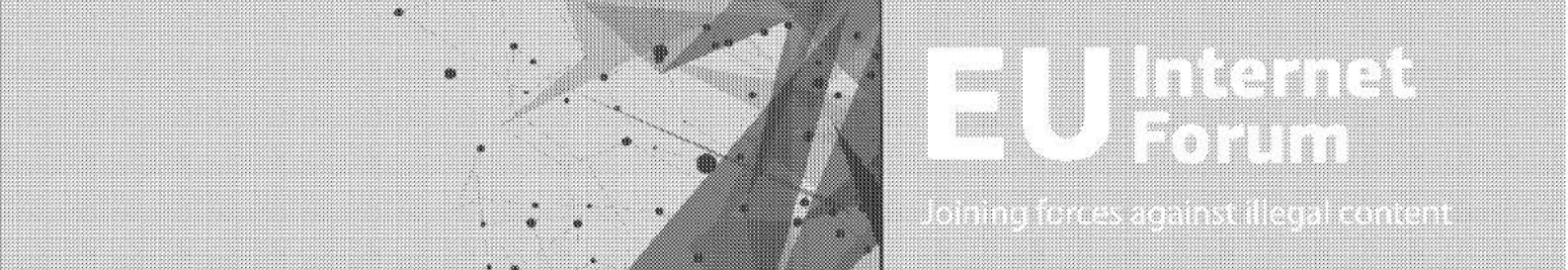
Not all flags necessitate review by company employees or moderators. Platforms like Reddit rely on community-based flagging to “upvote” or “downvote” comments that other users make on Reddit forums. These voting choices are used as signals: for other users, “upvotes” and “downvotes” can signify the quality of a post; for moderators (volunteer ones or company employees and moderators), these community-based votes can help in conducting a review process of a user, post, or forum.

Super User Flagging

Although all users may have the opportunity to flag content, the downside is that the quality of flagged content can be highly variable; many users do not necessarily select the appropriate flagging reason or flag content that they simply do not want to see. Other users, however, can be highly accurate in the content they flag and/or are highly active participants on a platform. In these instances, platforms may seek to incentivise this subset of users to report content by categorising them to be “moderators,” “trusted flaggers,” or “super users.” By differentiating the types of users, platforms can prioritise flags from super users over flags from the broader user base.

User-Reputation Based

This category is one that is in flux and is poorly discussed externally by platforms. Some companies incorporate user scores or trust scores to identify higher-risk users. Alternatively, users that have strong performance in identifying and reporting troubling content, irrespective of the level of activity they have on a given platform, can be seen as deserving a higher reputation than others. In this category, users who more regularly engage with a platform, or engage in ways that are less anonymous, may be given more features or tools, for example, prioritised flagging (see above), blocking privileges, etc. The key difference between “super user flagging” and “user-reputation based” is that the former is generally a determination made by company staff or law or regulation (e.g., the Digital Services Act), and the latter is a holistic determination of a user’s profile and general activity on the platform, not just the quality of the user’s flags.



	IN-PRODUCT FLAGGING	SUPER USER FLAGGING	USER-REPUTATION BASED
ADVANTAGES	Reflects an established norm across social media platforms that allows users to play a role in the content moderation enterprise.	Allows companies to identify certain users who have a strong record of finding violative content consistently.	Provides options for platforms that seek to restrict certain product features based on a user's reputation.
DISADVANTAGES	Users often report content that is not violative, generating significant "noise" for a company's content moderation efforts. User attitudes on what constitutes a violation does not always align with a platform's policies.	Creates tiers of users. Super users could also engender tensions between users on a platform.	Concerns of bias, harassing conduct where one user instructs others to harass or intimidate others, leading to poorer reputations.

Figure 4.6.2a – Reactive Manual Systems

This chart shows the several types of User Flagging used in Reactive Manual Systems

It is rare, however, for companies to send *all* flagged content for human review. Content might be flagged by users at rates or volumes that are cost-prohibitive for a company to review each one. Furthermore, flagged content also carries the risk of being imprecise; a user’s personal determination that content is “spam” may not meet a company’s own definition of this material.

Flagged content is often triaged through the use of rules, or heuristics, that company employees develop. These rules might prioritise or expedite human review of content flagged for more egregious policy violations (such as TVE content or child safety) over others, or for flagged content in some languages over others, depending on the linguistic skills of a reviewer workforce. Alternatively, a piece of content that has received many user flags may also be sent for review more urgently than content with only one or a handful of complaints. The range of rules at the disposal of a company is extensive, and a full inventory of the factors companies can use to develop rules is neither feasible due to a lack of transparency nor within the scope of this report.

Reviewers are able to subsequently examine the content and utilise a range of actions, from approving the material, age-restricting the content, recommending that the content should not be monetised, removing the content from a platform’s recommendation system, or removing the content entirely.

Discrete pieces of content are not the only items reviewed manually. User accounts, groups, or other entities on a platform can be reviewed by moderators as well.

MANUAL SYSTEMS



Proactive Manual Systems

Without waiting for external users to submit complaints ("flags"), company employees and contractors review content surfaced through, primarily keyword matches.



Reactive Manual Systems

The traditional format where content is sent for human review because of user flags or detected through automated systems.



Tool-Enabled Manual Systems

Emphasizes the internal tools a company uses to aid in the detection and enforcement of content.

Figure 4.6.2b – Types of Manual Systems

	PROACTIVE MANUAL SYSTEMS	REACTIVE MANUAL SYSTEMS	TOOLS-ENABLED MANUAL SYSTEMS
ADVANTAGES	Affords company staff flexibility to create permanent or situation-dependent keyword lists.	Provides an avenue for users to alert companies to bad content.	Beneficial aid for company staff to detect, prioritize, or even de-prioritize certain types of content to send for review.
DISADVANTAGES	Keyword lists are quite laborious, requiring significant upkeep. Malicious actors can more easily circumvent these.	Significant toll on human content moderators who review the vast majority of flagged/reported content, issues as well with scaling this approach in a sustainable, cost-effective manner.	Requires technical time and resources to build and maintain the tools necessary for manual systems.

Figure 4.6.2c – Manual Systems comparison

Automated Systems

Companies often tout dizzying numbers of removals, from millions to billions of pieces of content. To conduct this scale of content moderation, companies must rely on automated systems. Automated systems enable companies to identify content more aggressively and on much larger scales than manual approaches.

While manual systems emphasise the need for *human* intervention to *identify and review* potentially troubling material, automated systems are focused on the design, deployment, and maintenance of machine learning (ML) models. The ML aids are primarily categorised through supervised and unsupervised ML models.

Supervised ML

Companies create detection algorithms for a range of content, from TVE content to child safety to illegal goods. These algorithms are “trained” using the corpus of content that has already been reviewed by company staff, its content moderators, or through automated measures (e.g., auto-removals). Supervised ML models are “supervised” because this training data is labelled through taxonomies developed by company staff to better refine the types to be detected by the ML model. Supervised ML models require significant human involvement in labelling the training data accurately. Human reviewers carefully categorise content based on established guidelines, ensuring the model’s ability to distinguish between benign and extremist content.

For example, a company creating an algorithm to detect TVE content may engage in a labelling exercise in its historical corpus of removed TVE content to differentiate between particular actors, types of threats, or any number of desired markers that its staff sees fit. The more granular the labels, the better the algorithm can differentiate the types of content that warrant review and the types that can be automatically rejected.

Two advantages to supervised ML models are their precision and interpretability. Supervised models can be highly precise depending on the availability, depth, rigour, and accuracy of the labelled training data. Since these models are trained on data that are labelled based on a taxonomy and classification

system, then model behaviours and outcomes can be interpreted by companies or outside bodies. At the same time, the models are limited in terms of the quantity of data and the rigour of the labelling process. The labelling process requires human raters and can be quite expensive and even open to bias. Also, to be able to respond to new terrorist actors or TVE trends, supervised ML models might be slow to respond and require significant retraining and data relabelling.

Unsupervised ML

Unsupervised ML models do not use labelled training data. Instead, these models work by extracting patterns from unlabelled data, such as finding patterns through sets of images or text. Unsupervised ML models can also rely on techniques such as clustering to identify suspicious clusters that deviate from patterns the models find.

Unsupervised ML systems excel in their scalability and adaptability. Because there is no need to pre-label training data, these systems can handle more and new types of data without teams of raters to work through the training corpus. Additionally, because unsupervised ML systems operate by finding their own patterns from data, they can identify emerging variations in TVE content without prior knowledge or labelled data.

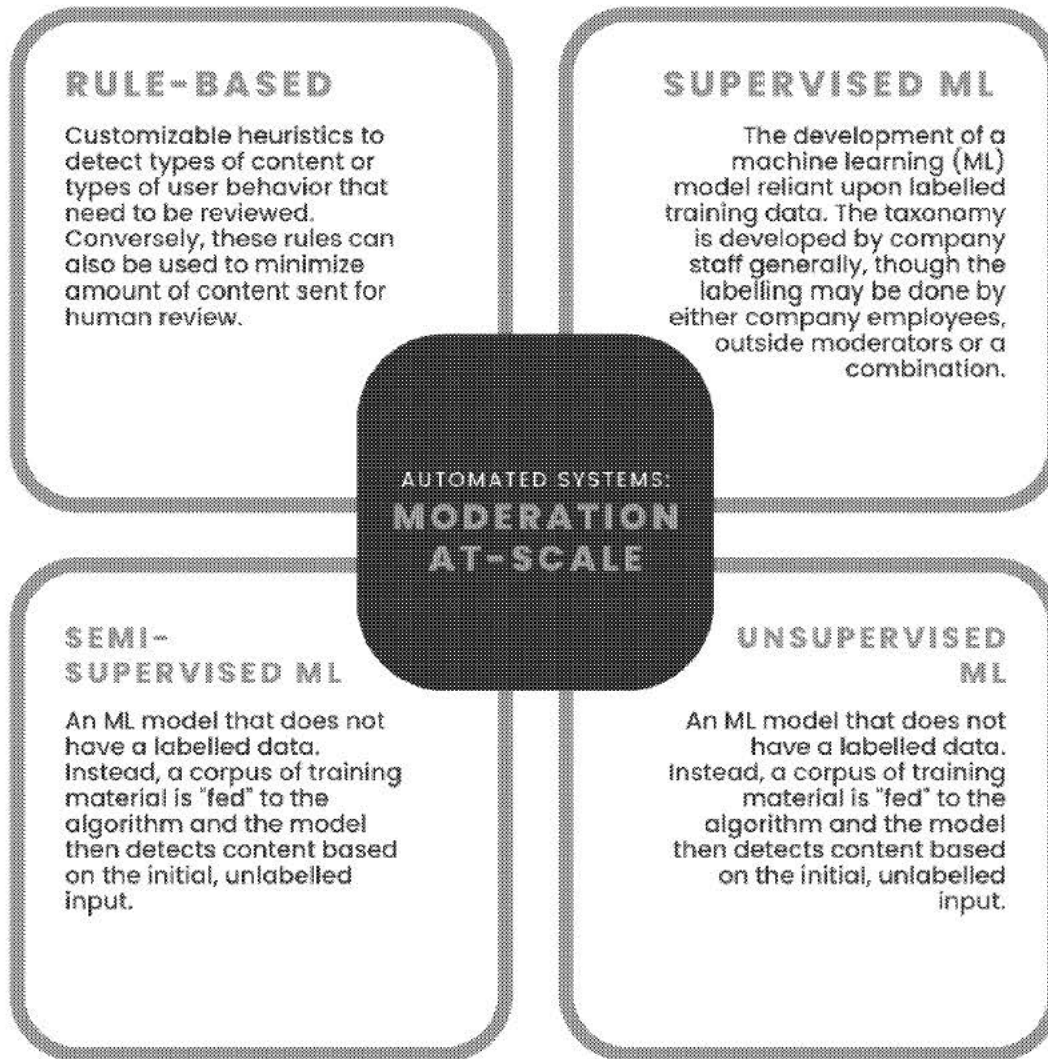


Figure 4.6.2d -- Types of Automated Systems

	RULE-BASED SYSTEMS	SUPERVISED ML	SEMI-SUPERVISED ML	UNSUPERVISED ML
ADVANTAGES	Fewer technical skills needed so more of a non-technical staff can create rules to detect, action, or enqueue content.	Can be highly tailored to particular policy areas or even particular trends or themes within a given policy vertical.	Can minimize the amount of company staff needed to create a robust taxonomy to label content.	No need to expose humans to graphic, controversial, or offensive content to label ML system.
DISADVANTAGES	Limited in what it can do. Oftentimes rules-based systems are best tailored to respond to specific, narrow issues.	Most expensive approach because of company staff time to develop taxonomy and the algorithm itself; human raters needed as well. Risk of bias as well depending on how content is labelled as well as the training data used.	See both supervised ML and unsupervised ML.	If an unsupervised ML was developed as a generalized algorithm to detect content across all policy areas, it might be face poor recall and precision. Risk of bias depending on the training data used.

Figure 4.6.2e - Automated Systems comparison

4.6.3 Tools

Tools underpin the development and success of the systems discussed above. Internal tooling allows a company to develop, say, a rules-based automated system or to develop the range of manual review queues on which manual systems rely. The details of internal tooling for any specific company are considered proprietary and confidential information, which makes assessment difficult.

External tools such as the ones developed by the Global Internet Forum to Combat Terrorism (GIFCT) are easier to assess. The GIFCT maintains a database of hashes -- essentially, digital fingerprints -- of known terrorist content, which enables member companies to both contribute violations discovered on their respective platforms and run the hashes in the database against their own corpora. GIFCT's database has historically focused on TVE content belonging to, related to, or produced by ISIS and al-Qaeda, and their affiliates. Because the contents of the database have been narrowly defined, the database offers high quality when analysed through the prisms of precision, recall, and consistency. ISIS and al-Qaeda have been closely tracked by subject matter experts and the organisations have clear markers of the content they produce, whether through iconography, media arms, or other indicators.

The GIFCT taxonomy has expanded recently to better address other forms of TVE content. The organisation's inclusion parameters now require that all hashes must be associated with one of the following: 1) the United Nations Security Council's Consolidated Sanctions List; 2) content that triggers the GIFCT's incident response protocol; and 3) content that is aligned with "behavioural inclusion parameters." As a result of the expansion, social media platforms can now share hashed content belonging to violent Right-Wing Extremist entities as long as the third prong of the GIFCT's taxonomy-behavioural inclusion parameters-are met. These parameters are: 1) The organisation cannot be a governmental entity; 2) There must be a violent extremist identifier (e.g., logo, code, iconography) to indicate affiliation with an organisation, group, movement, or ideology; 3) The

organisations must have a core hate-based ideology; and 4) The organisation must advocate for violence.

The GIFCT tool is also quite scalable and relatively low-cost. Because hashes are a cheap method, companies can easily share content through the API that provides connectivity to the GIFCT hash database.

Tech Against Terrorism (TAT) is another organisation working to combat online TVE content. TAT works independently and in tandem with GIFCT. Companies can choose to be members of both organisations, but in order to become a GIFCT member, one of the requirements is that platforms complete a mentorship program with TAT. TAT maintains its own knowledge platform for members and also offers bespoke product solutions for platforms that resemble third-party services, which are discussed in the following section.

Companies that are members of GIFCT or TAT are afforded immense discretion to choose how much they use these features, if at all. A company can use the GIFCT database, for example, only to pull hashes, only to contribute hashes or both. If a company uses the database to find similar material on their own platforms, the company has the capability to choose whether “hits” from the hash-sharing database will lead to automatic removal, normal review, or expedited review.

A smaller, more resource-constrained social media platform may use the GIFCT database to simply automatically block content it finds on its platform that matches a hash in the GIFCT database. In fact, the GIFCT database is particularly powerful for smaller platforms that may not have the financial or technical resources to build out a content abuse effort in the company’s early stages.

4.6.4 Third-party Services

Third-party (3P) companies have become an integral part of content moderation over the past decade. These companies and the services they offer can aid the largest and smallest platforms alike, from the development of detection algorithms to outsourced moderation teams to bespoke policy guidance and threat analysis. Unlike in-house solutions, where the cost of dealing with global abuse problems must be borne exclusively by a single company, 3P solutions allow the amortisation of technology costs, resulting in higher Return On Investment for pervasive abuse problems like TVE. Additionally, given the cross-platform nature of many abuse types, including TVE, complementary signals collected from many platforms result in a higher absolute performance than any one platform could achieve on its own. On the other hand, narrow yet impactful platform-specific abuse problems (such as gaming a particular bespoke feature) do not benefit from either of these advantages and are likely best handled in-house.

While there are many 3P providers in the Trust & Safety space, the services generally fall within the following categories:

- **Risk Intelligence & Measurement:** This typically entails combing through the open and dark web to detect trends and specific coordinated threats that company staff should be alerted to. In some cases, these trends and comparisons can directly benchmark platform performance and risk, creating actionable goals.
- **Lead Generation:** Often, a 3P will sift through a platform’s content to flag and report material that should have been previously detected and removed pursuant to a platform’s

community guidelines. They do this with experience (e.g., former government or military experts identifying TVE content based on specialised knowledge) and/or off-platform signals (e.g., relationships with known bad actors or forums, etc).

- **Content / User Filters:** Some 3P services provide filtering software that can block certain types of images, text, or video when provided access to a stream of content or user data. These filters generally work on an abuse topic basis (e.g., TVE, Spam, Hate Speech, etc), but some can also be generalised models. Some platforms may offer high quality yet narrow offerings, others offer 360 solutions with lower precision/recall.
- **Trust & Safety Platforms:** Some 3P companies provide an entire content moderation platform, providing companies an alternative to building internal tools and systems. These platforms can be tailored to create review queues, establish rules-based detection and enforcement actions, incorporate outside classifiers, and frequently have wellness features.

This space is rapidly evolving and, with the release of additional regulations, is likely to accelerate due to converging platform policies, obligatory transparency, and the standardisation of Trust and Safety requirements.

4.6.5 Content Moderation Recommendations

One of the best opportunities to improve TVE content moderation lies in creating more consistency in defining TVE across platforms at an example and policy playbook level. This standardisation will facilitate greater sharing of TVE or Borderline content in the social media ecosystem. Greater alignment also carries immense benefits in cost-savings to platforms as more consistent policies will ultimately aid 3P tools and services to meet the needs of multiple platforms simultaneously at a lower cost.

Platforms should leverage manual and automated systems that minimise the amount of content sent for human review by automating high-confidence TVE content. This can be done by leveraging external high-precision tools, such as the GIFCT hash-sharing database, and special flaggers, such as Trusted Flagger programs with NGOs or government agencies, due to the relatively low cost and time investment required to integrate these particular high-precision external TVE-fighting methods. Furthermore, platforms should rely on supervised ML systems that are trained on high-quality data from known violative samples to both scan on content upload (proactive automated systems) and after user flagging (reactive automated systems) as well.

To more effectively address TVE content that implicates and spreads across multiple platforms, companies should seek greater opportunities to work with, and consult the services of, 3P service providers. These partners can provide platforms with the opportunity to minimise the number of raters or in-house employees needed to develop systems or review content. In addition to cost savings, 3P service providers amortise R&D across many companies for industry-wide challenges. Given TVE's ever-changing nature—in terms of both threat actors and shifting government priorities—and the broader social media ecosystem where this content circulates, 3P service providers maintain a unique vantage point compared to any one platform focused on the confines of their own services. When working with these 3P service providers, however, the platforms should exercise caution to identify 3P service providers that understand the potential biases, limitations, and risks that arise with outsourced detection algorithms and Trust & Safety processes and have the knowledge and experience to mitigate these risks. Another additional advantage of 3P service providers is their independence, which removes any perception of bias or internal conflict of interest that a platform may have.

5 Conclusions and Recommendations

In this study, it has been established that all platforms are amplifying TVE content, as well as Borderline content through their established personalisation algorithms and recommender systems. It has also been demonstrated by platforms like TikTok that this amplification can be significantly less pronounced, despite having similar levels of TVE Findability as other platforms. For this reason, more transparency, measurement and knowledge sharing with regard to recommender system algorithms affecting TVE content among platforms, potentially facilitated by an independent third party, could lay the foundation for positive changes. This same level of transparency and collaboration can also benefit TVE Findability of harmful content, as demonstrated by YouTube (see Figure A.6.11.13). In this context, the provisions of the EU's Digital Services Act related to transparency and data access will require platforms to disclose information related to their content moderation practices and the functioning of their algorithms, providing key insights.

Platforms and academics should come together to further analyse the results of this study and propose additional mitigation measures to lessen the dissemination of TVE content. This large dataset, collected across the 5 social media platforms, 8 languages over a multi-month measurement time frame, has the potential to hold many more insights worth analysing, discussing and drawing new conclusions. In particular, it is important that academics, policymakers and social media platforms engage in earnest dialogue about policy, technical and regulatory implications of the findings, and what should be done to decrease the amplification of TVE content on social media. Trust Lab would be a keen participant in such efforts.

The most important recommendation that we are making is that there need to be additional, comprehensive and more frequent measurement studies like this one in the future. This study was limited in scope due to constrained resources considering the large number of platforms and languages to cover. As a result, statistical significance often suffered, and more detailed data deep dives to answer important secondary questions about dependencies within the data set could only be partially covered. In addition, the study limited the results to aggregate statistics and larger-than-desired confidence interval bands that didn't allow for an otherwise more differentiated analysis, ranking, or trend analysis.

More frequent measurement would also increase accountability by platforms to make improvements. It's not possible to know if social media platforms are reducing the amount of TVE amplification without rigorous measurements that assess the impact of platform and regulatory efforts in this area. Platforms, in particular, have an opportunity to move the measurement needle by providing the necessary data through such means as direct access to their internal systems, which would greatly reduce the expense and time in collecting outside-in data. A portion of the metrics in this study were not statistically significant because of time and budget constraints, amplified by the need to collect the data outside-in in a semi-manual fashion to adhere to platforms' terms of service. Also, repeated studies have the added benefit of measuring new effects that play into the social media ecosystem. For example, generative Artificial Intelligence (AI) is an emerging factor that will impact social media in many ways (both positively and negatively). We have also seen significant reductions in resource allocation towards platform safety, as well as significant individual governance and policy changes (e.g., Twitter) that have no doubt impacted TVE content on social media.

Differences in the type of TVE content has shown that the platforms are emphasising topics and content that media outlets and policymakers consider important. In doing so, platforms overlook or neglect certain other areas of interest (such as Left-Wing TVE content or Italian TVE content) and

focus most of their efforts on minimising the amount of International and Right-Wing TVE content that is being shown to users. Conversely, platforms are more effective when choosing to prioritise certain areas or are pressured to do so by regulators or media. For instance, the amount of International TVE content in Arabic language that can be found on platforms is low, and it's more often removed than other types of content.

The scope of this study did not include the assessment of how online amplification of TVE and Borderline content on social media platforms contributes to offline radicalisation among users, and the public in general. There are studies in this space looking at it from the perspective of the influencer rather than the victim²⁸. Further studies are necessary in the social science field to understand how exposure to such content via social media searches and feeds leads to changing of views and to radicalisation.

The authors of this report wish to commend the European Commission and the European Union Internet Forum for their leadership in working proactively with partners to stop terrorists from using the Internet to radicalise, recruit and incite individuals to violence. Special thanks to the EU Directorate-General Migration and Home Affairs for developing this important project. We are grateful for the opportunity to work with all of you and the five social media platforms on this project to protect EU citizens online.

²⁸ Thompson, R. (2011). Radicalization and the Use of Social Media. *Journal of Strategic Security*, 4(4), 167–190. <http://www.jstor.org/stable/26463917>
55

6 Appendix

6.1 Project Team

Table A.6.1 - Project Team Members

Name	Profile	Domain of Specialisation	Tasks	Location
Anna Maria CARPANI (Fincons)	Expert	6+ years with International Organisations and European Institutions. 20+ years in Project management.	Overall Project Management	Off-site (IT: Milan)
Tom SIEGEL (Trust Lab)	Expert	Trust & Safety for internet, organise event and speech. Content safety, privacy and security protections for most of Google's products including Websearch, YouTube, Play, Social Products, Ads, Payments, Cloud, Gmail and many others	Task 3, Task 6	Off-site (US: Palo Alto / DE: Berlin)
Ray LIU (Trust Lab)	Expert	Director of Trust & Safety at Google for 15 years. Leading global policy and enforcement teams across different products, including Ads, Publisher and Developer products.	Task 3, Task 6	Off-site (US: Bay Area)
Peter DUDIČ (Trust Lab)	Expert	Trust & Safety Manager at YouTube for 6+ years. Expert knowledge of social media hate speech law NetzDG and T&S policy verticals.	Task 1, Task 2, Task 4	Off-site (SK - Trnava)
Nicholas MILLER (Trust Lab)	Expert	Data Scientist, experienced in massive data analysis. Oversees key areas such as methodology, tooling, and data across global measurement projects. Multi-national corporations' data problems.	Task 1, Task 2, Task 3, Task 4	Off-site (JP: Tokyo/ UK: London)
Fabienne MEIJER (Trust Lab)	Expert	Consultant with a focus on human-centred innovation with clients ranging from Fortune 500 companies to start-ups. 7+ years in data and research work.	Task 1 Task 2 Task 3 Task 4	Off-site
Benji LONEY (Trust Lab)	Expert	Co-founder of Trust Lab. 10+ years in Trust and Safety YouTube, TikTok and Reddit experience.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site (US: Whitefish)
Shankar PONNEKANTI (Trust Lab)	Expert	Co-founder of Trust Lab. Distinguished Engineer with 15+ years of experience.	Task 1 Task 2 Task 3 Task 4	Off-site (US: Bay Area)

			Task 6	
Amre METWALLY (Trust Lab)	Expert	Graduate of Harvard Law School and has 5 years of work experience in digital rights, freedom of expression, and online safety work. Previously responsible for policy and enforcement issues at YouTube related to political extremism, counterterrorism, and graphic violence. Amre has also published extensively on technology and human rights issues in several academic law journals.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Fara ISLAM (Trust Lab)	Expert	Cyber Safety and Research Analyst at Trust Lab. Accelerated Masters of Public Policy Candidate at the University of Virginia with a keen interest in cyber ethics and responsible technology. Fara interned at TikTok as a Global Issue Policy Intern covering Harassment and Bullying on the platform, while exposing herself to the different verticals such as Violent Extremism, Graphic Content, and Integrity and Authenticity.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Theodoros EVGENIOU (Tremau)	Expert	Chief Innovation Officer. 25+ years experience in Machine Learning. Professor, INSEAD.	Task 5	Off-site (FR: Paris)
Ian William CHRISTENSEN (Tremau)	Expert	Trust & Safety / TVE Data Scientist and Policy Analyst at Tremau. Bachelor thesis in environmental data analysis at Columbia University.	Task 5	Off-site (DE: Berlin)
Ra MOUR (Tremau)	Expert	AI research at Tremau. Bachelor in Computation and Cognition from MIT.	Task 5	Off-site (US: Cambridge)
Xuqin WANG (Tremau)	Expert	AI researcher at Tremau. Phd Student in Computer Vision at TUM.	Task 5	Off-site (DE: Munich)

6.2 Keywords

These are the lists of keywords that were used during the Search Phases. These keywords were distributed in English and localised by native speakers for market appropriateness.

Table A.6.2a - International terrorism (Islamic terrorism - ISIS, al-Qaeda and its affiliates)

Keyword	Type of keyword	Comments
Anwar al Awlaki	Actor/Agent	Leader in al-Qaeda
Abu Yahya al-Libi	Actor/Agent	High-ranking official within al-Qaeda
Abdullah Azzam	Actor/Agent	Teacher and mentor to Osama bin Laden, al-Qaeda related
Ayman al-Zawahiri	Actor/Agent	Prominent leader of al-Qaeda
Shehzad Tanweer	Actor/Agent	One of four al-Qaeda-related terrorists in London bombings
Hizb ut-Tahrir	Group/Org	Terrorist group linked to violent acts in multiple countries.
Ansar al-Khilafah	Group/Org	Philippine-based militant group pledging allegiance to ISIS
Amaq Agency	Group/Org	News outlet that is first point of publication for claims of responsibility for ISIS and al-Qaeda attacks
Hezbollah	Group/Org	Lebanon-based terrorist group involved in numerous anti-US terrorist attacks
Muslim Brotherhood	Group/Org	Debate on recognising this as foreign terrorist organisation (FTO). Minor links to violent tactics in prior years, but currently rejects the use of violence and seeks to be more involved in the mainstream political process.
Soldiers of the Caliphate	Phrase/Slogan	Gunmen who have carried out terrorist attacks
How to do Jihad	Phrase/Slogan	The term "Jihad" here is used in extremism context, where terrorists recruit people for al-Qaeda and ISIS-related activities
Killing Kuffar	Phrase/Slogan	Extremist ideal of killing a disbeliever of al-Qaeda and ISIS-related ideology
suicide bomber	Phrase/Slogan	When an individual sacrifices their life to commit a terrorist attack
Dabiq Magazine	Phrase/Slogan	Online magazine used for al-Qaeda and ISIS-related recruitment and announcement
#islamicterrorism	Hashtags	Terrorism related to al-Qaeda and ISIS
#DespitetheDisbelievers	Hashtags	Refers to disbelievers of al-Qaeda and ISIS-related ideology
#IslamicState (#ISIS, #ISIL)	Hashtags	A name for al-Qaeda and ISIS
#caliphatestate	Hashtags	A name for al-Qaeda and ISIS
#Beheadings	Hashtags	Signature method for al-Qaeda and ISIS-related terrorism. It refers to the removal of the head from the body.

Table A.6.2b - Violent Right-Wing Extremism (far-right, extreme nationalistic groups & movements, neo-nazis)

Keyword	Type of keyword	Comments
Richard Spencer	Actor/Agent	American neo-Nazi, antisemitic conspiracy theorist, and white supremacist.
David Duke	Actor/Agent	American white supremacist, antisemitic conspiracy theorist, far-right politician, convicted felon, and former leader of the Ku Klux Klan.
Nikolai Nerling	Actor/Agent	German Right-Wing Extremist, anti-Semite and Holocaust denier. Alias name is "Der Volkslehrer".
David Lane	Actor/Agent	American domestic terrorist, white separatist, neo-Nazi, convicted felon, and important ideologist of white supremacy.
Olena Semenyaka	Actor/Agent	Figurehead of the Azov movement and Ukrainian Nationalism
Blood and Honour	Group/Org	International neo-Nazi/racist skinhead group started by British white supremacist
Atomwaffen Division	Group/Org	international far Right-Wing Extremist and neo-Nazi terrorist network
Nordic Resistance Movement	Group/Org	Pan-Nordic neo-Nazi movement in the Nordic countries and a political party in Sweden
Kolovrat	Group/Org	Far-right groups in a number of Slavic countries, mainly in Russia and Ukraine, use the Kolovrat in place of a swastika. It was also included in the emblem of the Russian Right-Wing radical movement Russian National Unity.
Prussian Blue	Group/Org	Prussian Blue was an American white power music duo formed in early 2003 by their mother April Gaede, a member of the neo-Nazi organization National Vanguard. Their music was described as racist and white supremacist, promoting neo-Nazi rhetoric such as Holocaust denial.
rapefugees	Phrase/Slogan	Blend of words rape and refugee.
Odinism	Phrase/Slogan	Ideology of Germanic paganism
1488	Phrase/Slogan	White Supremacist Hate Symbol. More context: The first symbol is 14, which is shorthand for the "14 Words" slogan: "We must secure the existence of our people and a future for white children." The second is 88, which stands for "Heil Hitler" (H being the 8th letter of the alphabet).
white power accelerationism	Phrase/Slogan	Idea that is inspiring white supremacist killers around the world; promotion of white supremacy
Diversity is a code word for <u>white genocide</u>	Phrase/Slogan	"Diversity = White Genocide" is a white supremacist slogan
#whitegenocide	Hashtagsf	White supremacy Ideology that white supremacy is dying
#refugeesnotwelcomed	Hashtags	White supremacy idea that refugees do not belong.

#pegida	Hashtags	Patriotic Europeans Against the Islamicisation of the Occident, abbreviated Pegida, is a pan-European, anti-Islam, far-right political movement.
#istandwithrussia	Hashtags	Showcases alliance with Russia
#whitelivesmatter	Hashtags	Promotion of white supremacy

Table A.6.2c - Violent Left-Wing Extremism (anarchism, communism, etc.)

Keyword	Type of keyword	Comments
Pierre Carette	Actor/Agent	Leader of radicalised Left-Wing group (Belgium)
Otelo Saraiva de Carvalho	Actor/Agent	Leader of radicalised Left-Wing group (Portugal) - deceased
Andreas Baader	Actor/Agent	RAF terrorist
Gudrun Ensslin	Actor/Agent	RAF terrorist
Communist Combatant Cells (CCC)	Group/Org	Belgian
The Revolutionary People's Liberation Party/Front	Group/Org	Known radicalised Left-Wing group
Organisation Anarchist Action	Group/Org	Known radicalised Left-Wing group
International Revolutionary Front	Group/Org	Known radicalised Left-Wing group
Red Army Faction	Group/Org	Communist, anti-imperialist, and urban guerrilla group engaged in armed resistance against what they deemed to be a fascist state
Red Brigades	Group/Org	Militant Left-Wing organisation in Italy linked to violent acts
Lone actors	Phrase/Slogan	Violent anarchist terrorists
We shall force them to their knees	Phrase/Slogan	Antifa phrase
extreme anarcho-socialists	Phrase/Slogan	Individuals that promote violence with anarchism
Antifascist	Phrase/Slogan	A member of ANTIFA or anyone against facism
Eat the rich	Phrase/Slogan	ANTIFA slogan against capitalism
#capitalistmedia	Hashtags	Term to say that media only follows capitalist ideas and rejects anarchism
#anarchism	Hashtags	Political philosophy and movement that is sceptical of all justifications for authority and seeks to abolish the institutions
#pan-destroyer	Hashtags	Associating violence with Marxism
#antifa	Hashtags	Left-Wing anti-fascist and anti-racist political movement in the United States
#acab	Hashtags	Left-Wing acronym for All Coppers Are Bastards

6.3 External Experts

Trust Lab partnered with a number of highly qualified EU and internal experts during the data collection and labelling phase of the project, to ensure exceptional data quality and integrity.

For the data labelling process, we worked with academic TVEC experts who were also native speakers in one or several of the researched languages. As well as industry practitioners with experience in TVEC analysis and moderation, who were also proficient in the researched languages. Both groups (academic and industry), independent from each other, labelled the collected data according to our provided standards.

Trust Lab Internally, Amre Metwally, a TVEC expert and lawyer who worked at YouTube and Clubhouse for many years, oversaw the quality process. He performed regular quality checks on the work done by external parties and routinely provided feedback to improve existing processes.

6.4 Task 1: Findability Charts

6.4.1 Findability per Platform

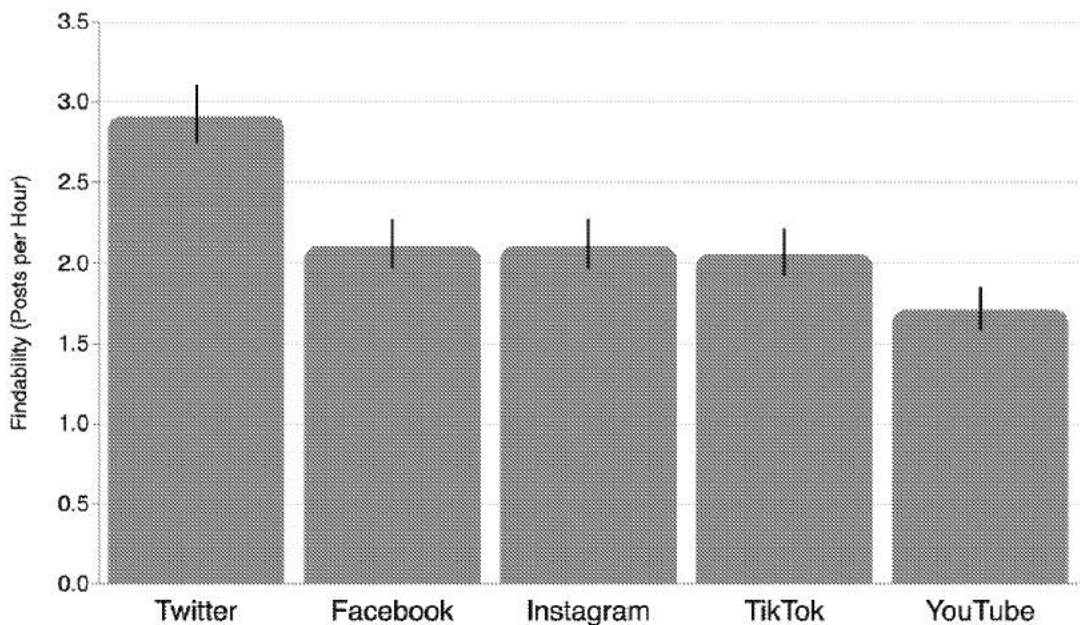


Figure A.6.4.1

The Findability Score is the average amount of content a motivated user can find on a given platform in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0012	0.0014	0.0030	0.0000
Twitter	0.0000	0.0000	0.0000	
TikTok	0.7036	0.7137		
Instagram	0.9930			

Table A.6.4.1

P-values on Findability per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.4.2 Findability per Language

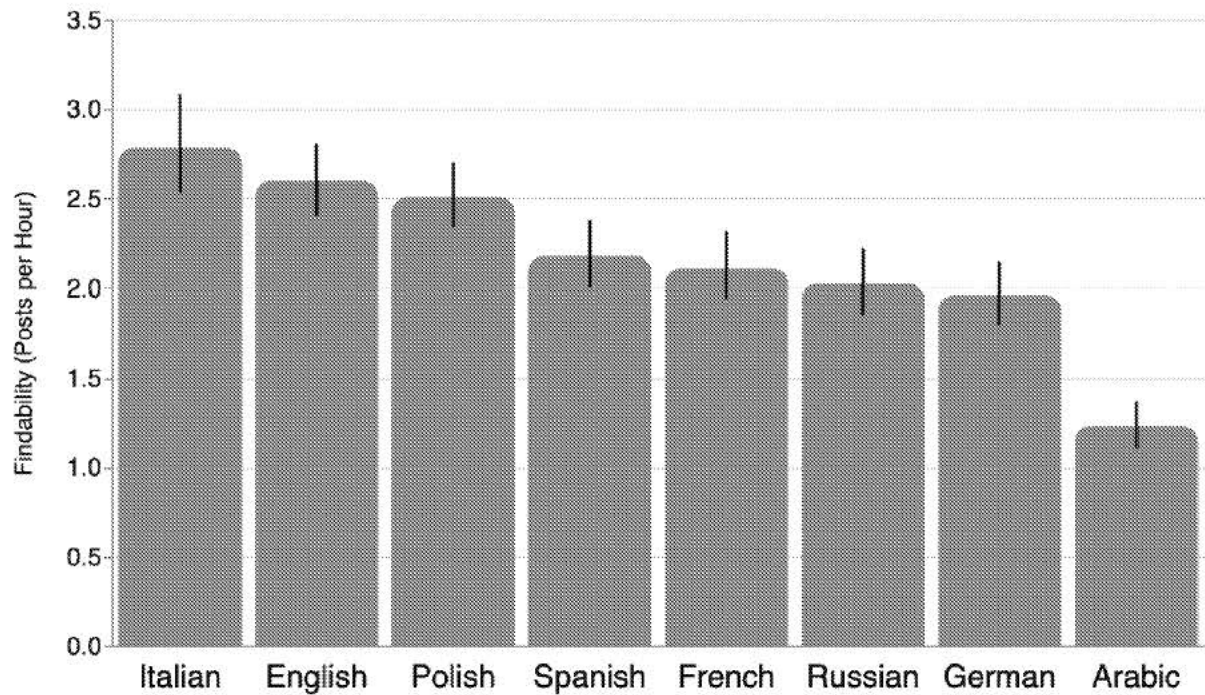


Figure A.6.4.2

The Findability Score is the average amount of content a motivated user can find in a given language in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0000	0.0115	0.6674	0.1624	0.0029	0.0348	0.3312
Russian	0.0000	0.0006	0.5919	0.6856	0.0002	0.0019	
Polish	0.0000	0.5823	0.0113	0.0003	0.1694		
Italian	0.0000	0.3710	0.0009	0.0000			
German	0.0000	0.0001	0.3426				
French	0.0000	0.0034					
English	0.0000						

Table A.6.4.2

P-values on Findability per Language

$\alpha = 0.00625$ (Bonferroni correction from 0.05)

6.4.3 Findability per TVE Type

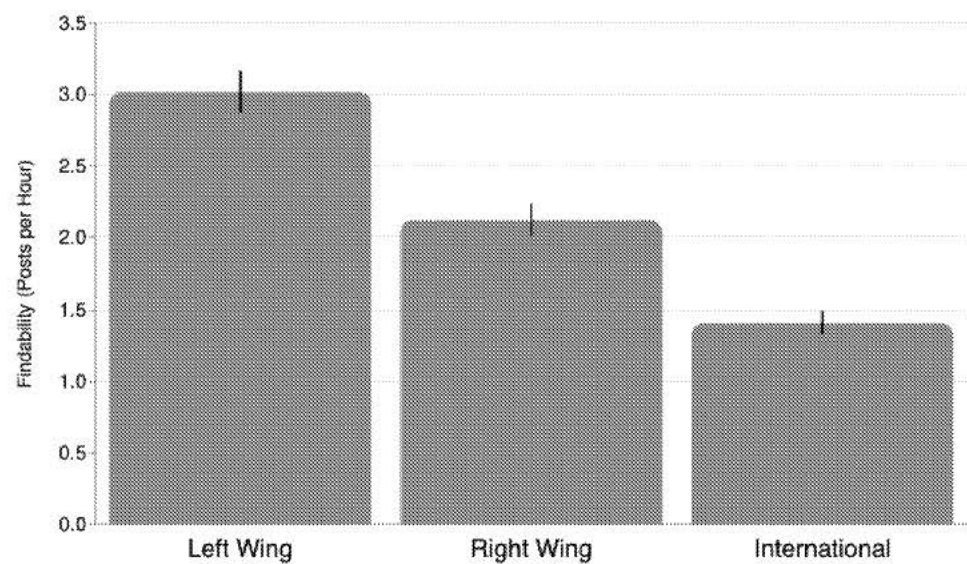


Figure A.6.4.3

The Findability Score is the average amount of content a motivated user can find for each TVE type in a one-hour search period. The block lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0000	0.0000
Left Wing	0.0000	

Table A.6.4.3

P-values on Findability per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.4.4 Findability per Language / TVE Type

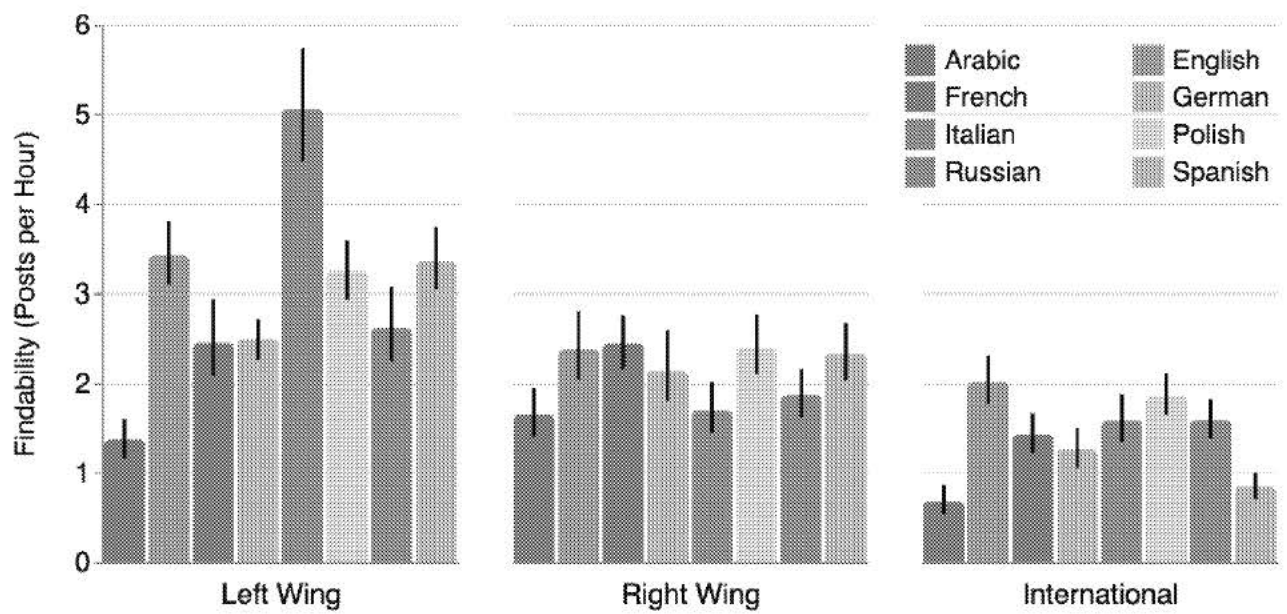


Figure A.6.4.4

The Findability Score is the average amount of content a motivated user can find for the respective language and TVE type in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

6.5 Task 1: Examples of Italian Violent Left-Wing Extremist Content

To help illustrate some of the Violent Left-Wing Extremist content, here is an example of a relatively benign [video](#) on YouTube that supports anarchism:



Image A.6.5a

The video title translates to “What if you were also an Anarchist?” and the bio of the user reads (translated using Google Translate):

“Anyone who hears the term Anarchy thinks of chaos, disorder, violence. Family upbringing and society have inculcated a totally false and distorted view of Anarchy into people’s minds. Try doing this kind of test and find out if, like me, you are an Anarchist.”

- A more extreme example is below (warning: violent), where the poster of the Tweet has the name **Redacted** and includes the hashtag ACAB (All Cops are Bastards) accompanying a photo of uniformed police officers, one who is surrounded by fire. The name of the user **Redacted** or the hashtag alone, without any other signal, might indicate Borderline content by indicating association, or at the very least affinity for, a Left-Wing organisation (Antifa) or a Left-Wing ideology (anti-police). However, with the image, the content, in the eyes of the outside expert, would constitute a violation for encouraging the use of violence:

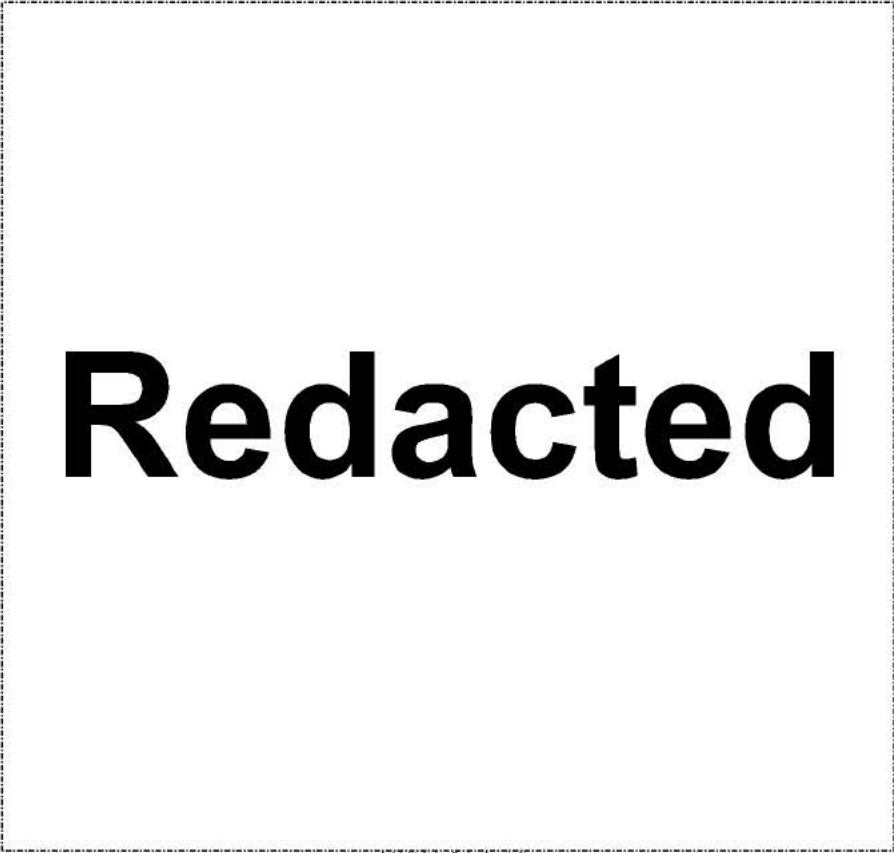
Redacted

6.6 Task 1: Examples Twitter Content

Some examples of content that was discovered on Twitter. This section starts with Right-Wing (Borderline) content, then Left-Wing (Borderline) content, then concludes with International (Borderline) content.

6.6.1 Violent Right-Wing Extremism and Borderline Examples

- Borderline anti-immigrant and anti-refugee rhetoric:



Redacted

Image A.5.6.10

- “White lives matter” rhetoric from anti-immigrant voices in Europe:

Redacted

Image A.6.8.1b

- Praise for **Redacted** and other leading figures and voices in the white supremacist movement. The user’s bio purports that the account is intended to be a parody. While many

platforms may have exceptions for satirical or comedic purposes in certain instances, they become difficult to enforce. One key reason is that discrete pieces of content often lack broader context, especially if platform review processes separate user-level review from content review. If so, then a reviewer simply looking at this screenshot would see clear praise for Richard Spencer based on the text. If reviewers are able to see this piece of content along with the user bio that states it is a parody account, the analysis becomes more complicated.



Redacted

Image A6.6.1c

6.6.2 Violent Left-Wing Extremism and Borderline Examples

- Critiques of capitalism and capitalist structures (e.g., media):

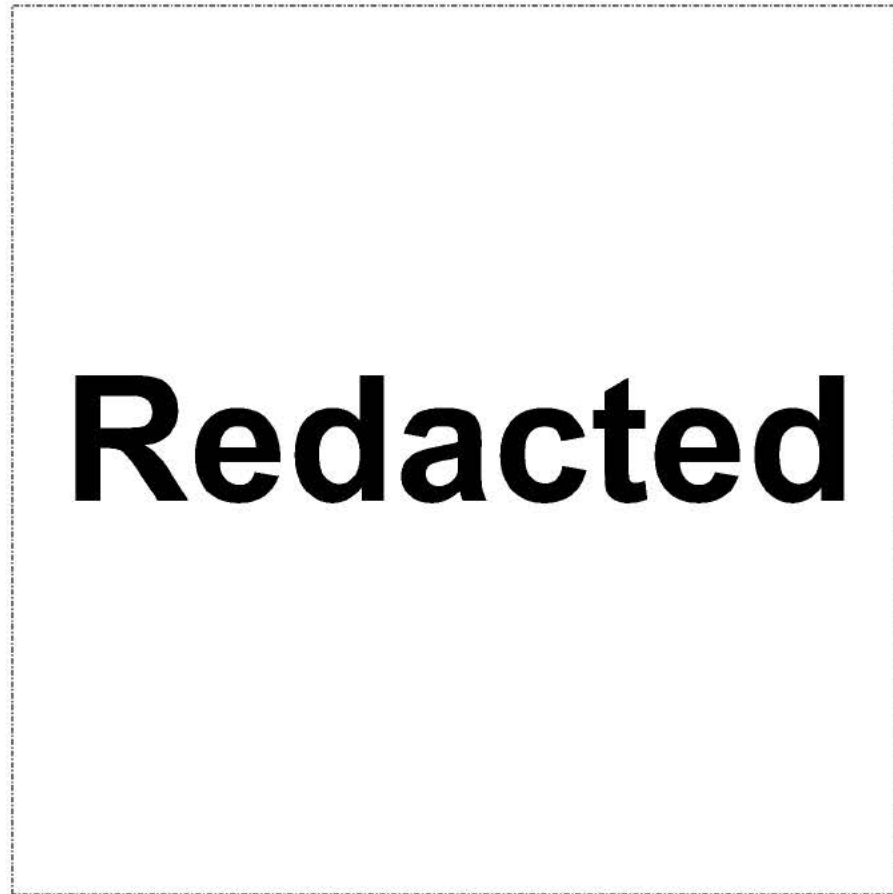


Image A.6.6.2a

- Violent Left-Wing Extremist content exemplified through anarchist rhetoric :

Redacted

image A.6.6.20

- Anti-police commentary:



Redacted

Image A.6.6.2c

6.6.3 International Extremism and Borderline Examples

Of the content related to Islamist TVE, it seemed to focus primarily on ISIS, Hezbollah, al-Qaeda, and Hizb-ut-Tahrir material. Examples include:

- Accounts dedicated to Hizb-ut-Tahrir “branches” in specific countries. The text says: “Independence of the flag from the colonialists”.

Redacted

- Praise for al-Qaeda leaders:

Redacted

Image A.6.5.3b

- ISIS-produced propaganda:

Redacted

Image A.6.6.3c

6.7 Task 1: User Engagement Charts and P-Values

6.7.1 Average Number of Shares

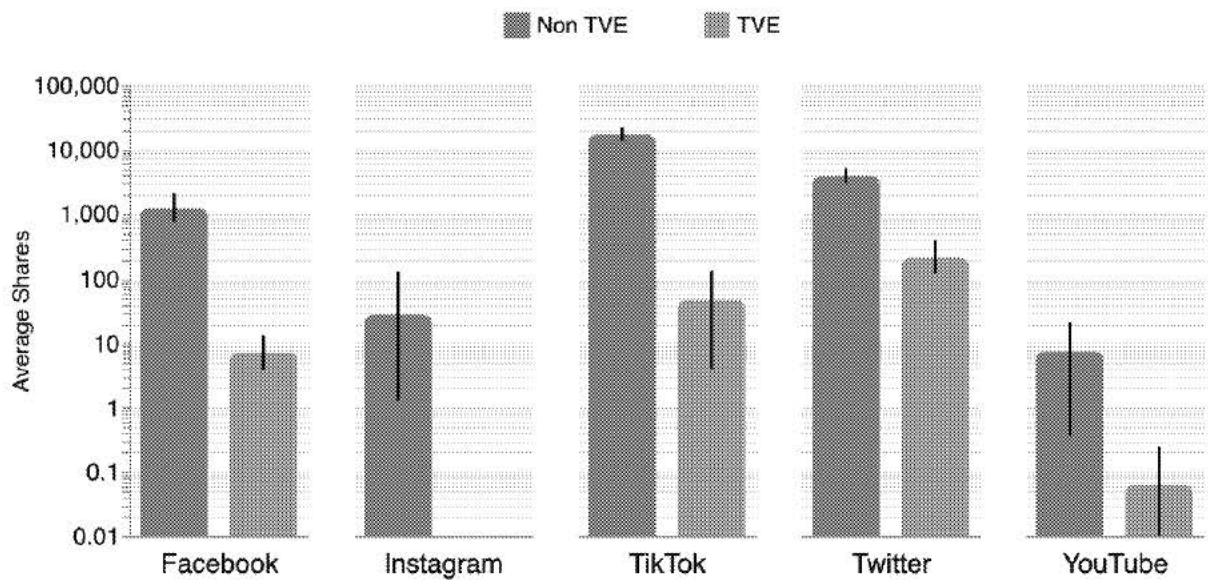


Figure A.6.7.1

This chart shows the average number of shares for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.²⁹

²⁹ Since the range of values in these charts is very high, it would be difficult to visualise the bars that have lower values. Taking the logarithm makes the values closer in range on the log scale and hence easier to visualise the chart as a whole. Therefore, for the interaction charts, we are using log scales.

6.7.2 Average Number of Likes

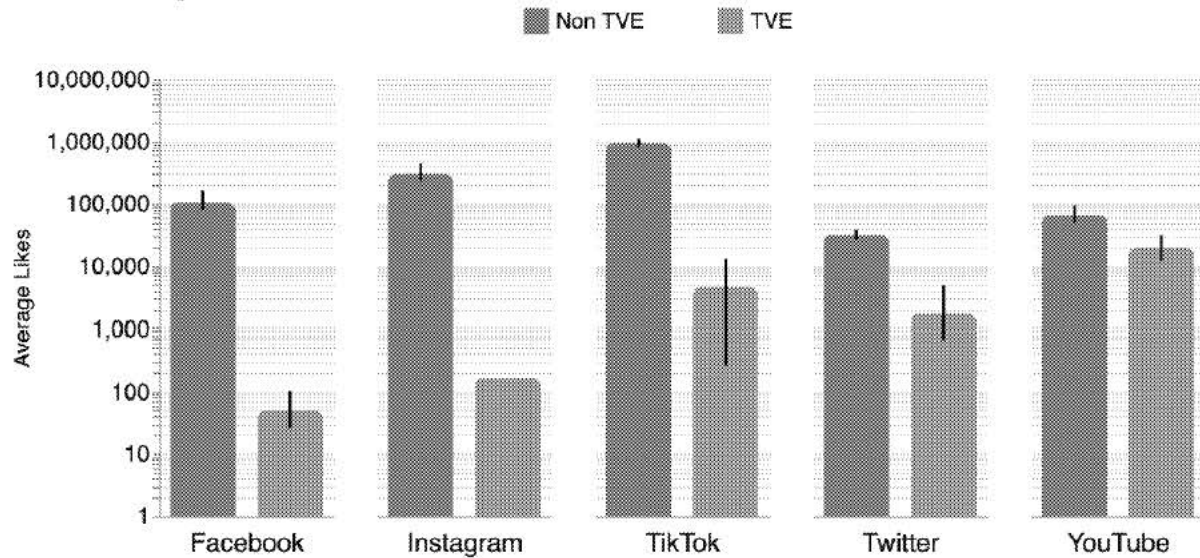


Figure A.6.7.2

This chart shows the average number of likes for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.

6.7.3 Average Number of Comments

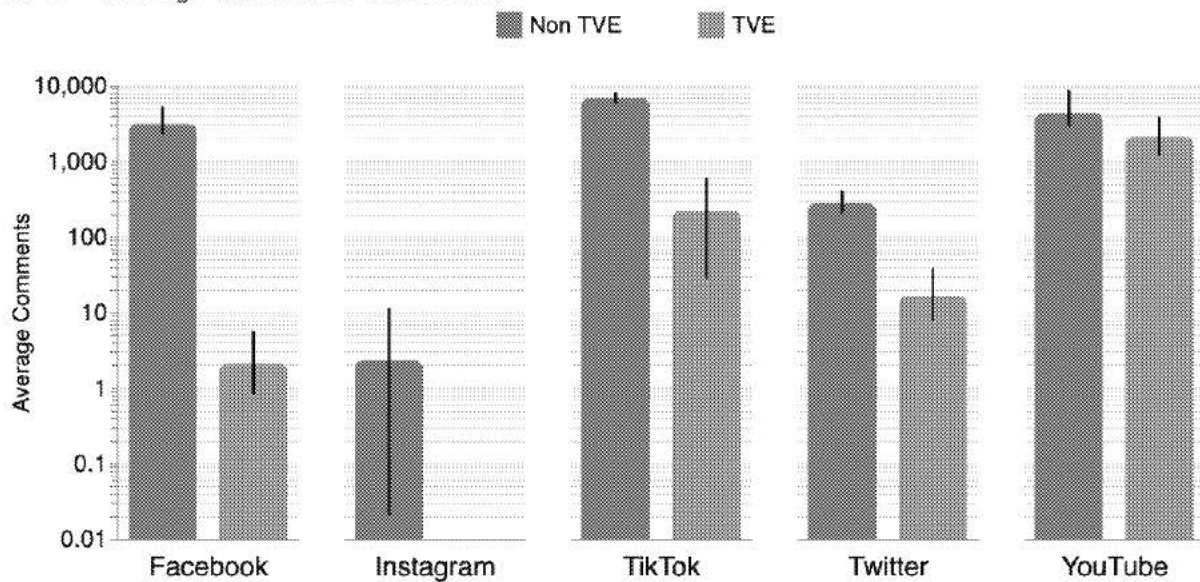


Figure A.6.7.3

This chart shows the average number of comments for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.

6.7.4 Average Number of Followers

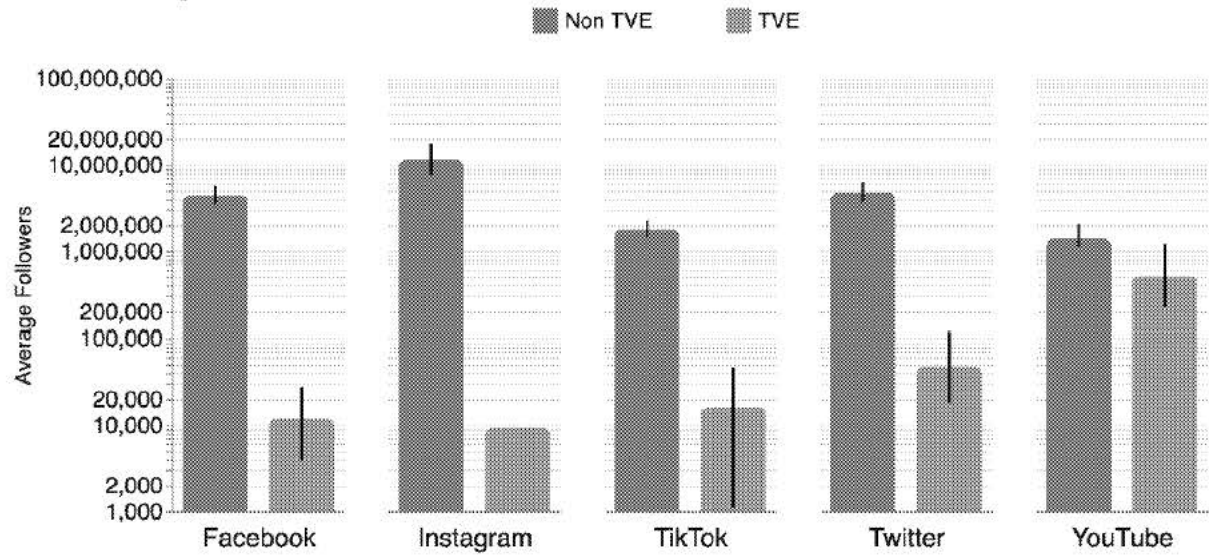


Figure A.6.7.4

This chart shows the average number of followers for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.

6.8 Task 1: Removal Rates for TVE Content

6.8.1 Removal Rates per Platform

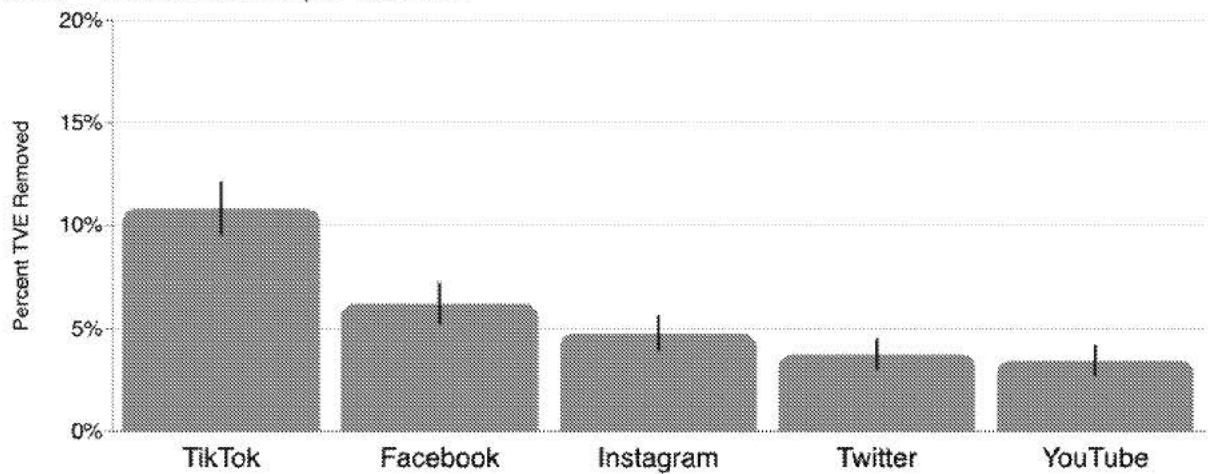


Figure A.6.8.1
This graph shows the percentage of TVE content that was removed on a given platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0311	0.2628	0.0000	0.7646
Twitter	0.0497	0.3870	0.0000	
TikTok	0.0062	0.0002		
Instagram	0.2802			

Table A.6.8.1
P-values on Removal Rates per Platform
alpha = 0.01 (Bonferroni correction from 0.05)

6.8.2 Removal Rates per Language

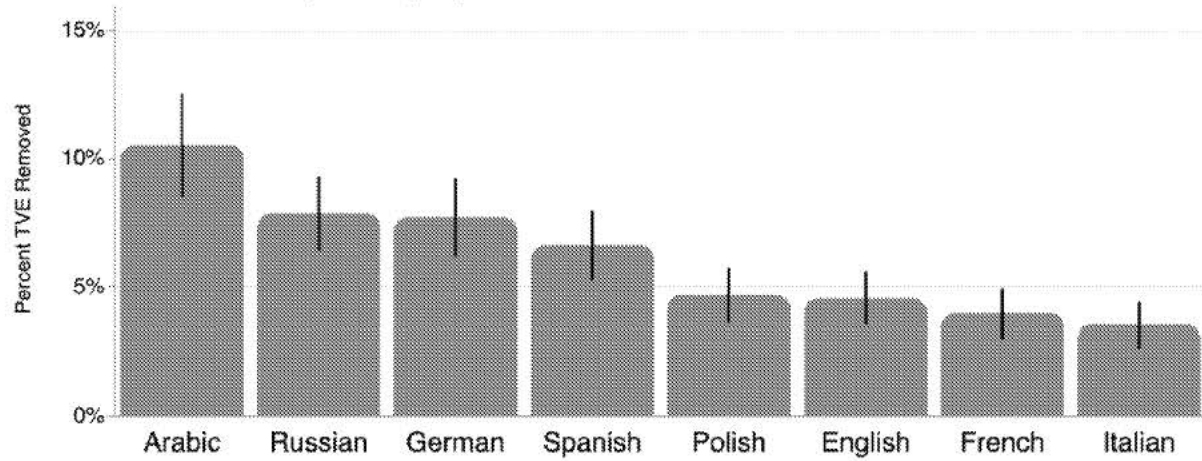


Figure A.6.8.2

This graph shows the percentage of TVE content that was removed in a given language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0957	0.2222	0.1053	0.5905	0.0490	0.2588	0.5354
Russian	0.2750	0.0618	0.0234	0.9472	0.0086	0.0766	
Polish	0.0053	0.9317	0.6224	0.0965	0.4042		
Italian	0.0003	0.4522	0.7349	0.0124			
German	0.2606	0.0791	0.0316				
French	0.0011	0.6825					
English	0.0039						

Table A.6.8.2

P-values on Removal Rates per Language

$\alpha = 0.00625$ (Bonferroni correction from 0.05)

6.8.3 Removal Rates per TVE Type

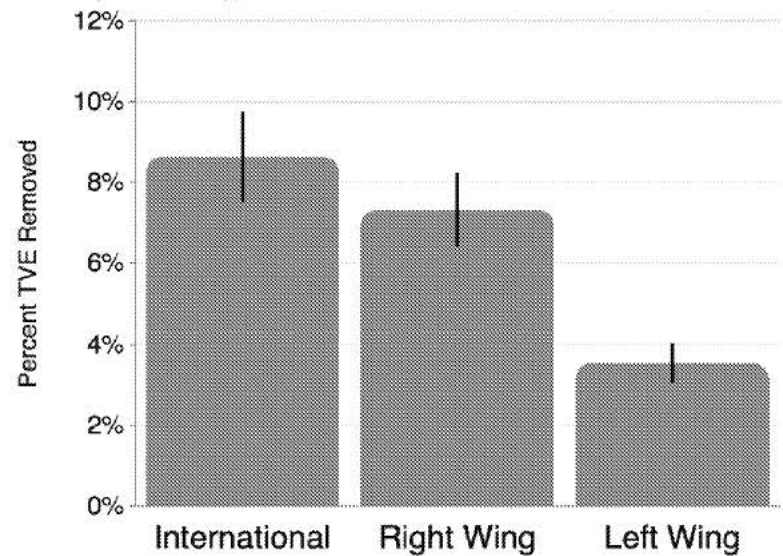


Figure A.6.8.3

This graph shows the percentage of TVE content that was removed per TVE type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.3610	0.0001
Left Wing	0.0000	

Table A.6.8.3

P-values on Removal Rates per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.8.4 Average number of Shares associated with Removed TVE

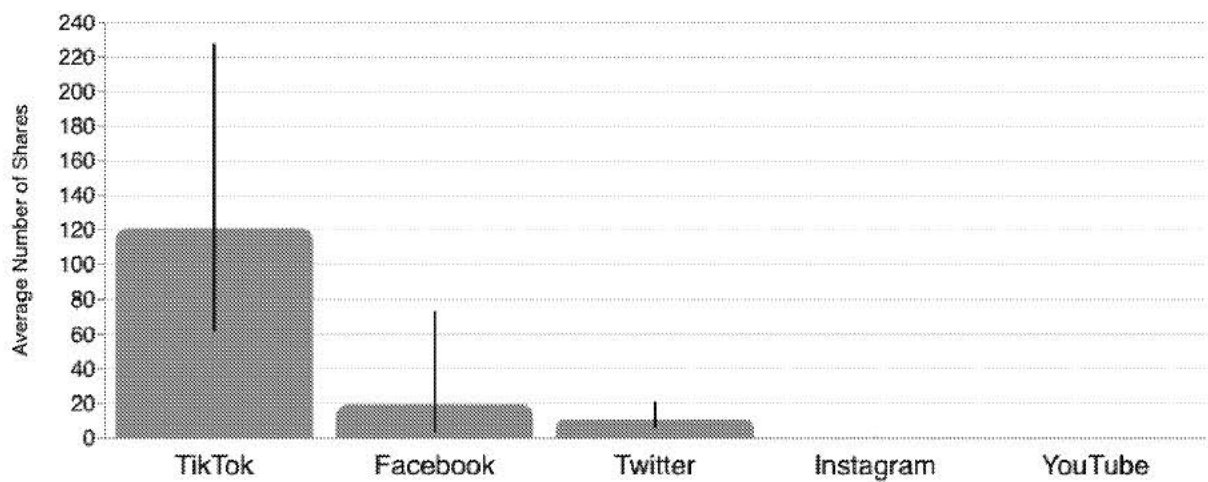


Figure A.6.8.4

This is the average number of shares associated with TVE content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.3076	0.2936	0.1426	0.0386
Twitter	0.5656	0.0160	0.1352	
TikTok	0.1001	0.0781		
Instagram	0.2784			

Table A.6.8.4

P-values on Average number of Shares associated with Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.5 Average number of Shares associated with TVE that wasn't Removed

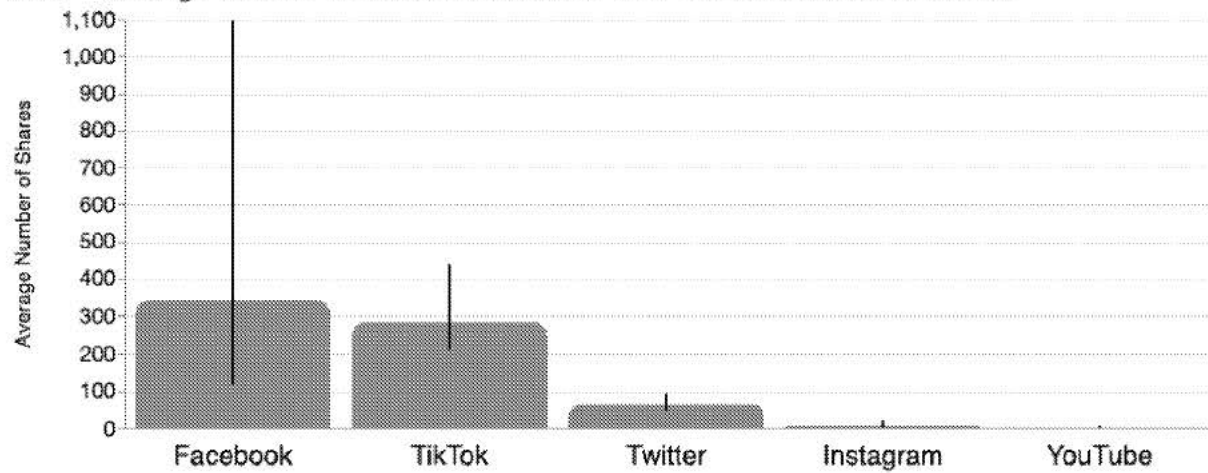


Figure A.6.8.5

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study. The block lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1245	0.3010	0.0000	0.0000
Twitter	0.1781	0.0001	0.0001	
TikTok	0.8029	0.0000		
Instagram	0.1198			

Table A.6.8.5

P-values on Average number of Shares associated with non-Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.6 Average number of Shares associated with TVE that wasn't Removed broken down by language

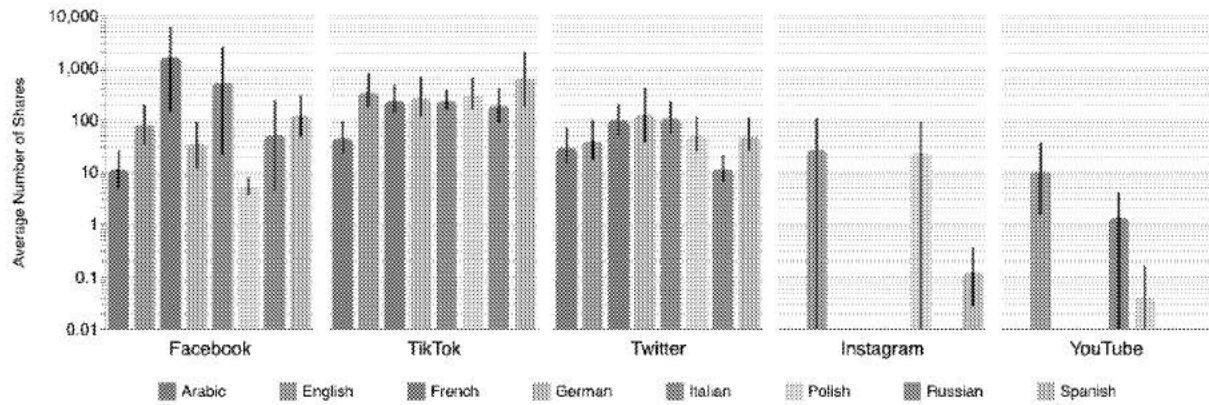


Figure A.6.8.6

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.8.7 Average number of Shares associated with Removed Borderline content

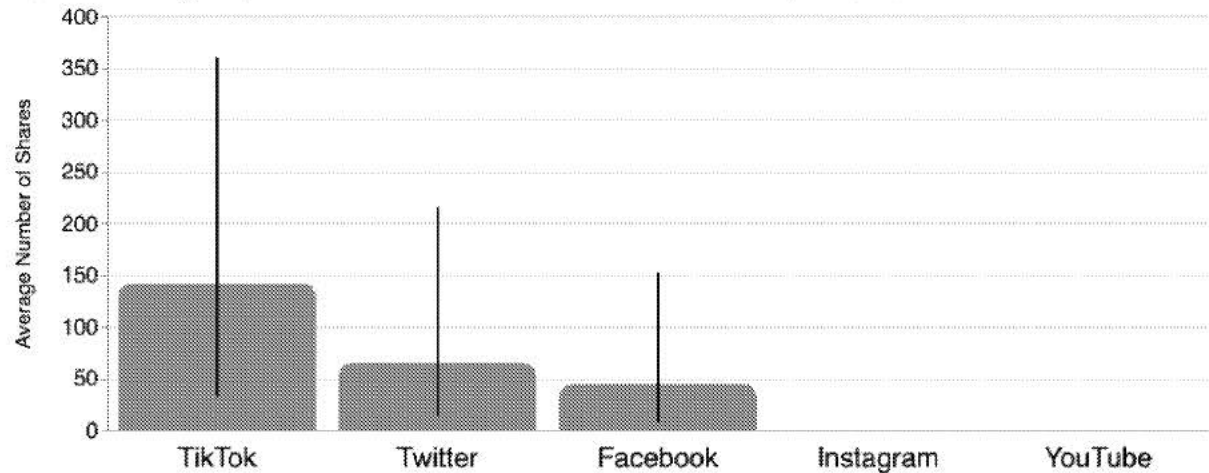


Figure A.6.8.7

This is the average number of shares associated with Borderline content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1305	0.0000	0.1387	0.1215
Twitter	0.6911	0.1194	0.4868	
TikTok	0.3310	0.1567		
Instagram	0.1596			

Table A.6.8.7

*P-values on Average number of Shares associated with Removed Borderline content
alpha = 0.01 (Bonferroni correction from 0.05)*

6.8.8 Average number of Shares associated with Borderline content that wasn't Removed

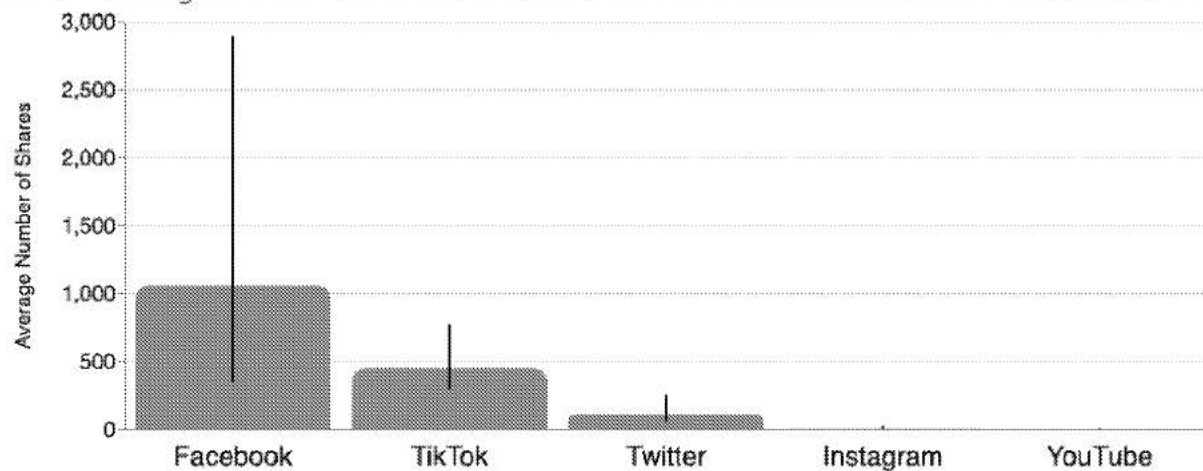


Figure A.6.8.8

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0617	0.5061	0.0001	0.0149
Twitter	0.1141	0.0302	0.0083	
TikTok	0.3682	0.0002		
Instagram	0.0861			

Table A.6.8.8

*P-values on Average number of Shares associated with non-Removed Borderline content
alpha = 0.01 (Bonferroni correction from 0.05)*

6.8.9 Average number of Shares associated with Borderline content that wasn't Removed broken down by language

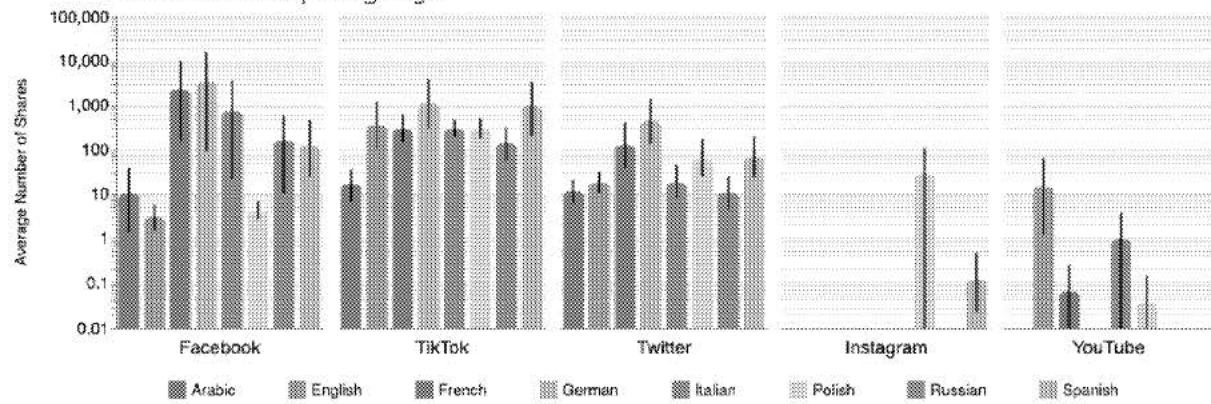


Figure A.6.8.9

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.9 Task 1: Removal Time

6.9.1 Removal Time per Platform

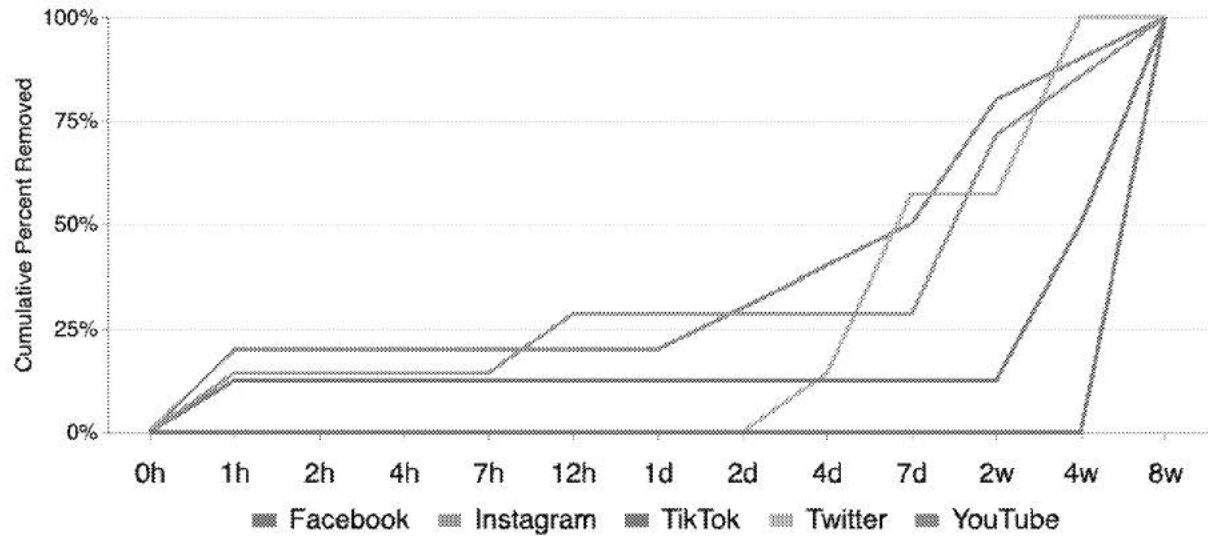


Figure A.6.9.1

This chart shows how long it took for each platform to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.2 Removal Time per Language

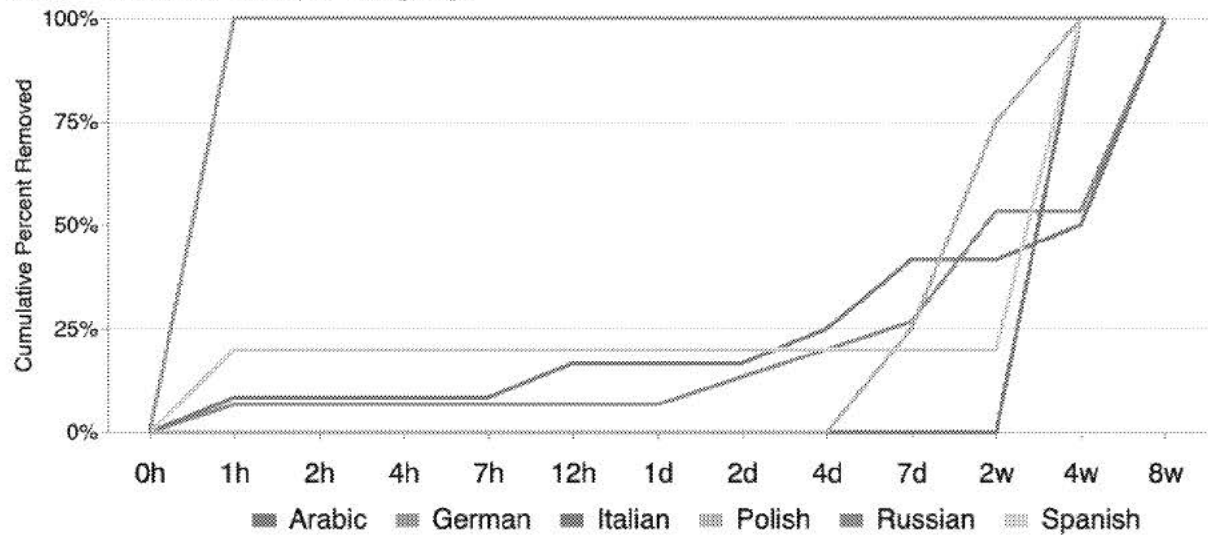


Figure A.6.9.2

This chart shows how long it took for each language to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.3 Removal Time per TVE Type

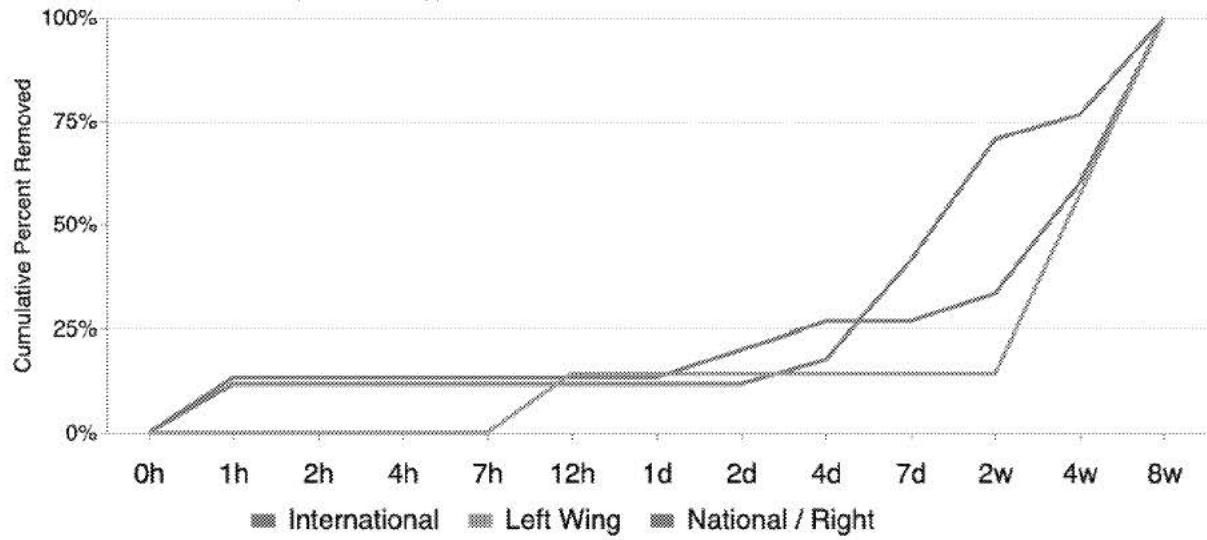


Figure A.6.9.3

This chart shows how long it took for each TVE Type to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.10 Task 1: User Sentiment Metrics

6.10.1 Severity Ratings by Platform

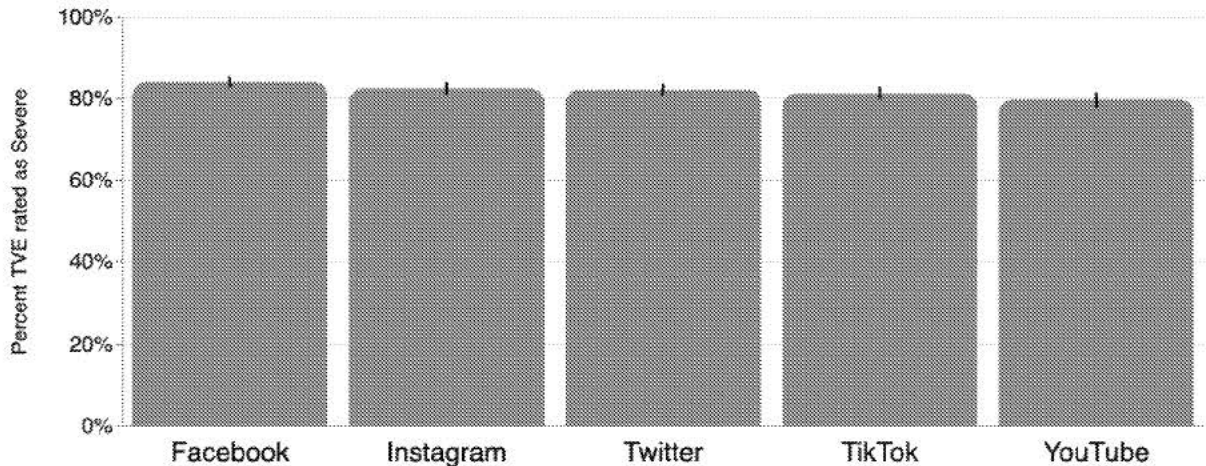


Figure A.6.10.1

This chart shows the percentage of TVE content present on different platforms that was rated by users as “severe”. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.10.1

P-values on Severity Ratings per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.10.2 Severity Ratings by Language

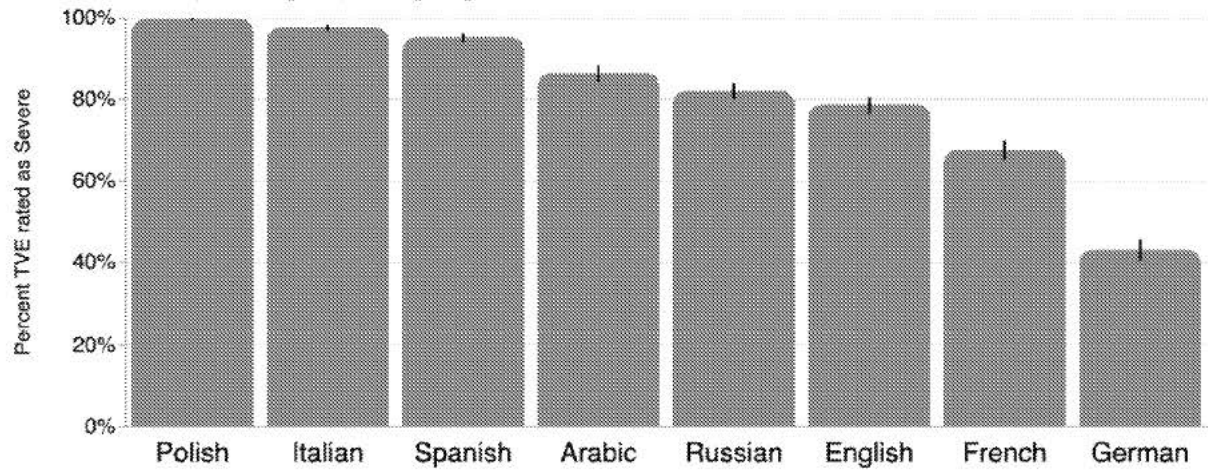


Figure A.6.10.2

This chart shows the percentage of TVE content present in different languages that was rated by users as severe. The black lines in the bars represent 90% confidence intervals.

Polish values are biased here because of the presence of bad actors in the dataset, which caused users to rate everything in the dataset as “severe”. This problem was investigated and deemed not present for the other languages.

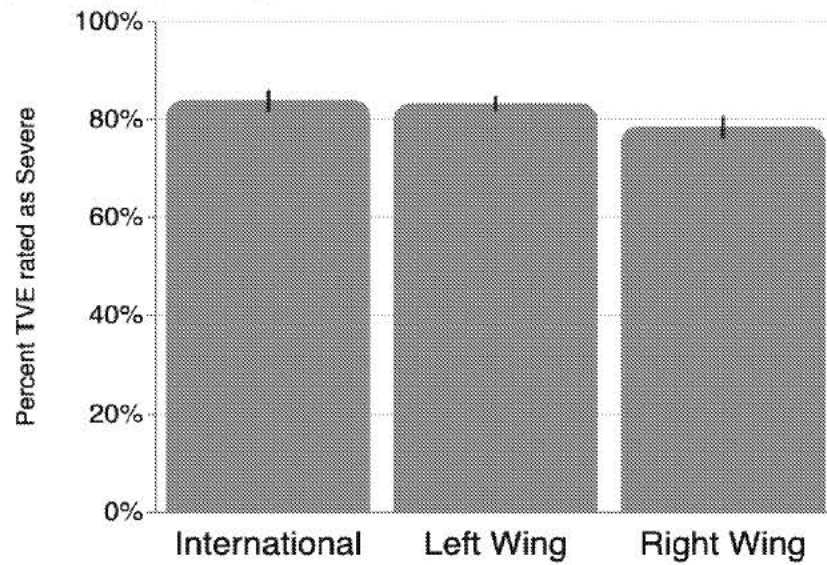
	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0001	0.0000	0.0000	0.0000	0.0573	0.0000	0.0000
Russian	0.1311	0.2206	0.0000	0.0000	0.0000	0.0000	
Polish	0.0000	0.0000	0.0000	0.0000	0.0124		
Italian	0.0000	0.0000	0.0000	0.0000			
German	0.0000	0.0000	0.0000				
French	0.0000	0.0003					
English	0.0088						

Table A.6.10.2

P-values on Severity Ratings per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.10.3 Severity Ratings by TVE Type



This chart shows the percentage of TVE content present for different TVE Types that was rated by users as "severe". The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0074	0.0045
Left Wing	0.6996	

Table A.6.10.3

P-values on Severity Ratings per TVE Type
alpha = 0.017 (Bonferroni correction from 0.05)

6.10.4 Severity Ratings over Time per Platform

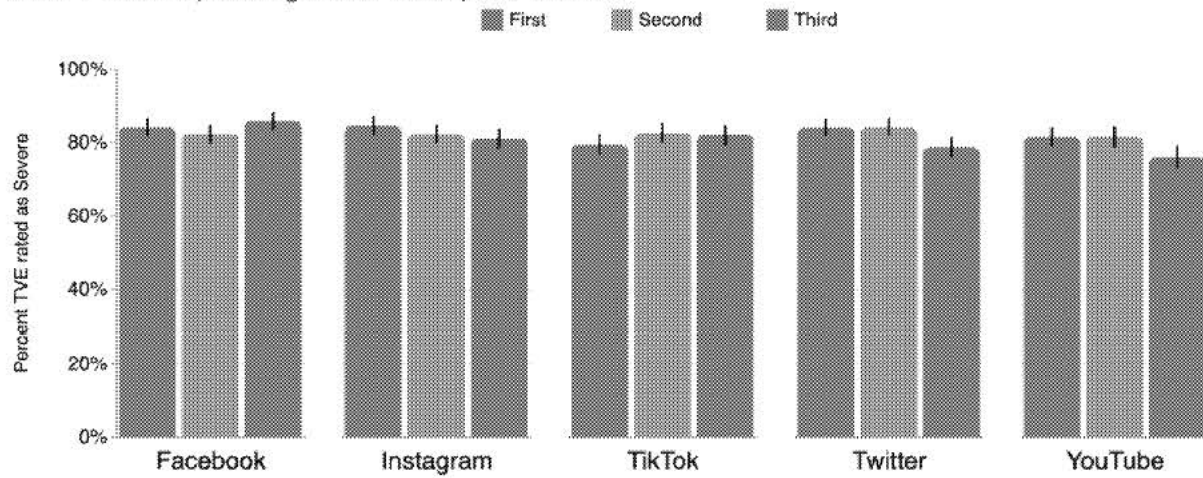


Figure A6.10.4

This chart shows the percentage of TVE content rated as "Severe". First, second and third stages are shown for each platform. The black lines in the bars represent 90% confidence intervals.

P-values are not included for this chart due to the number of combinations and complexity of the comparison.

6.11 Task 2: Amplification Charts

6.11.1 Amplification Across all Platforms

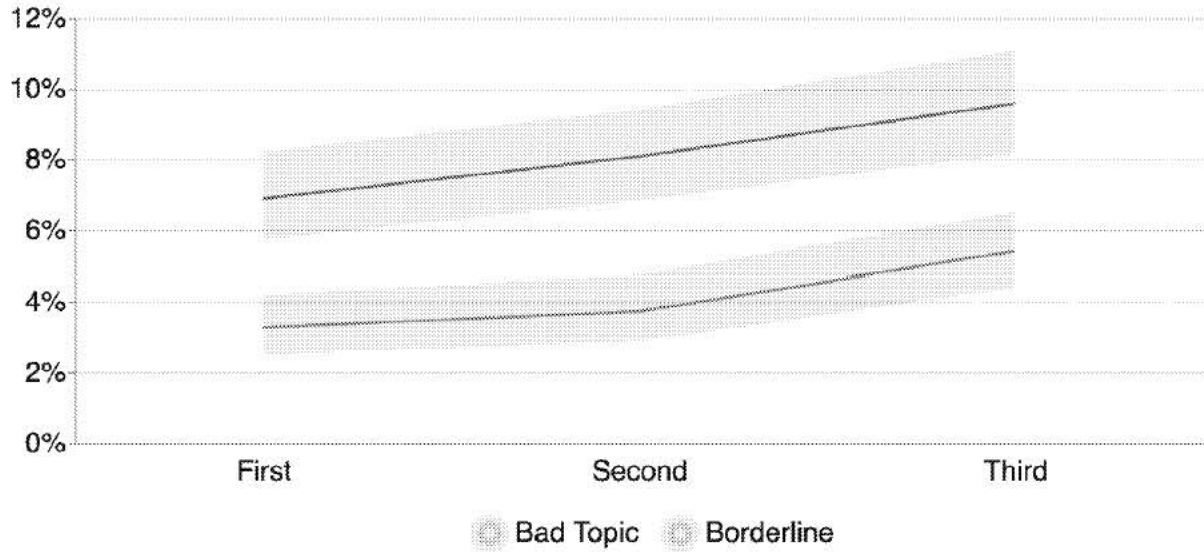


Figure A.6.11.1a

This graph illustrates the mean percentage of "Bad Topic" and "Borderline" content recommended in user feeds over time. The shaded area represents the 90% confidence interval ($\alpha = 0.05$)

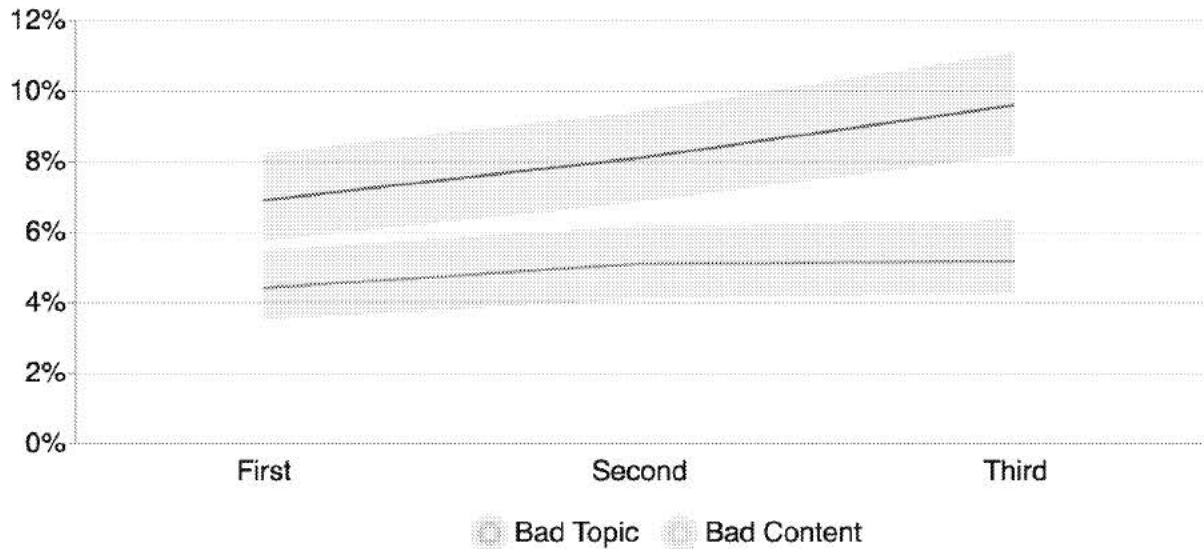


Figure A.6.11.1b

This graph illustrates the mean percentage of "Bad Topic" and "Bad Content" content recommended in user feeds over time. The shaded area represents the 90% confidence interval ($\alpha = 0.05$)

6.11.2 Amplification Average Percent of Bad Content in Feed per Platform

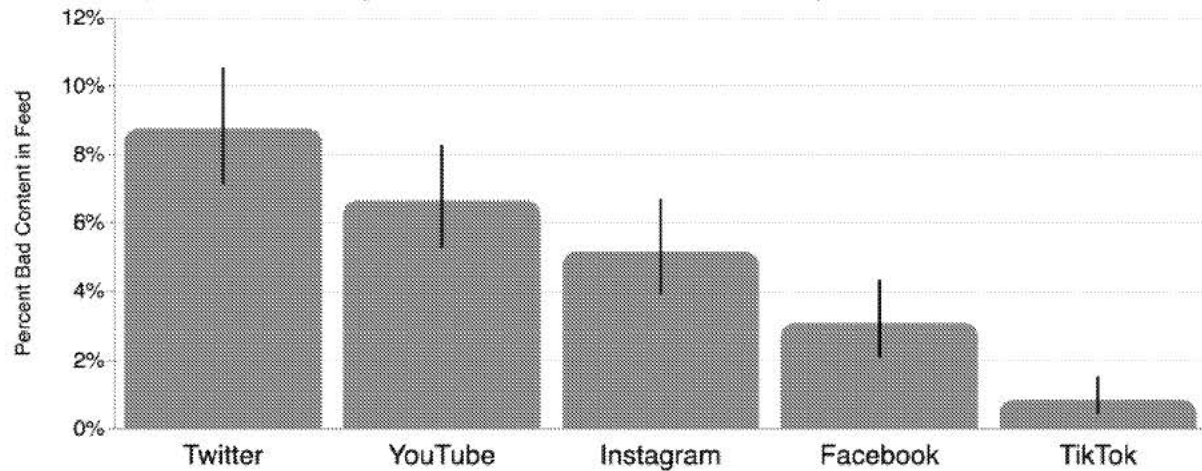


Figure A.6.11.2

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective platforms. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.11.2

P-values on Amplification Average Percent of Bad Content in Feed per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.11.3 Amplification Average Percent of Bad Content in Feed per Language

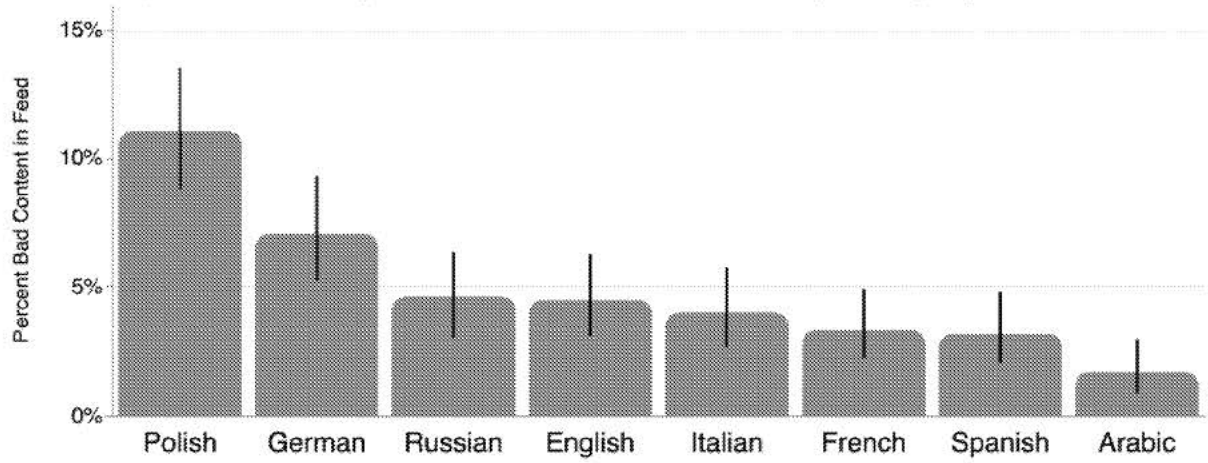


Figure A.6.11.3

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective languages. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.1466	0.3265	0.9038	0.0097	0.5335	0.0000	0.2809
Russian	0.0118	0.9246	0.3343	0.1168	0.6438	0.0003	
Polish	0.0000	0.0002	0.0000	0.0387	0.0001		
Italian	0.0381	0.7145	0.6132	0.0440			
German	0.0001	0.0986	0.0126				
French	0.1146	0.3855					
English	0.0155						

Table A.6.11.3

P-values on Amplification Average Percentage of Bad Content in Feed per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.11.4 Amplification Average Percent of Bad Content in Feed per TVE Type

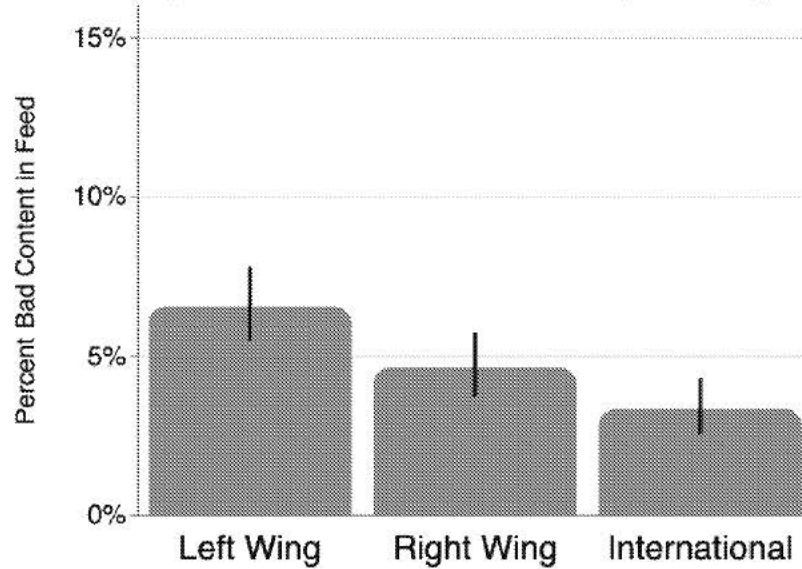


Figure A.6.11.4

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective TVE Types. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.1058	0.0442
Left Wing	0.0003	

Table A.6.11.4

P-values on Amplification Average Percent of Bad Content in Feed per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.11.5 Amplification Percentage Change for Bad Content per Content Type

To better compare data across other dimensions (platform, language, etc), we will use the percent difference between the first and third stage's mean. We don't compare to zero state because the assumption is that a new account's feed will be empty or contain zero TVE related content.

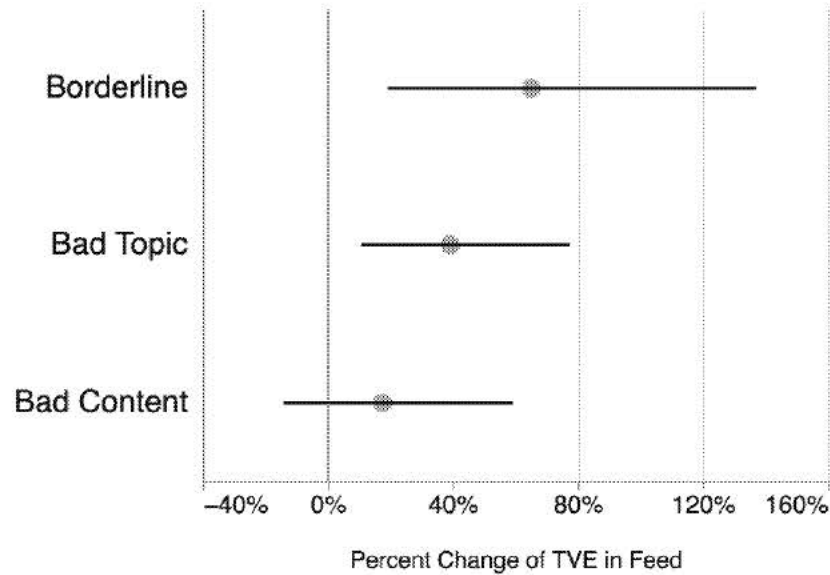


Figure A.6.11.5

This chart shows the percentage change for the amount of Bad (TVE) Content per Content Type in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.6 Amplification Percent Change for Bad Content per Platform

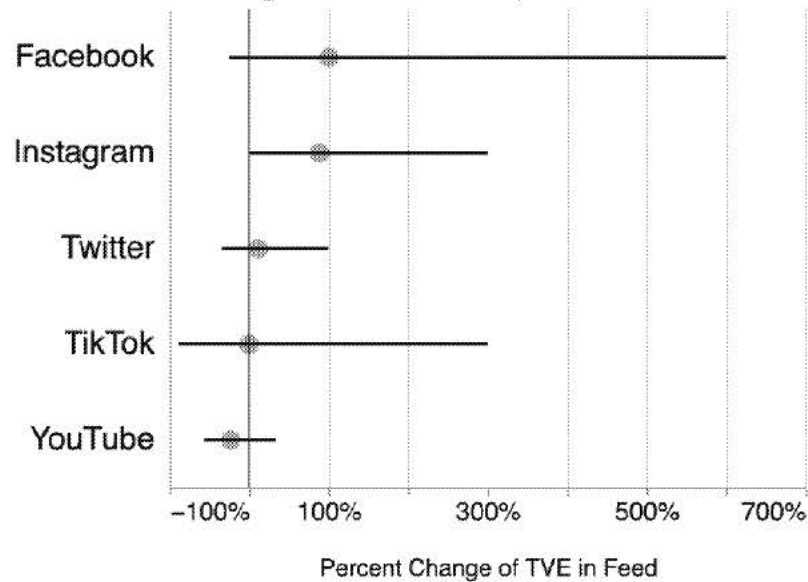


Figure A.6.11.6

This chart shows the percentage change for the amount of Bad (TVE) Content in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.7 Amplification Percent Change for Bad Content per Language

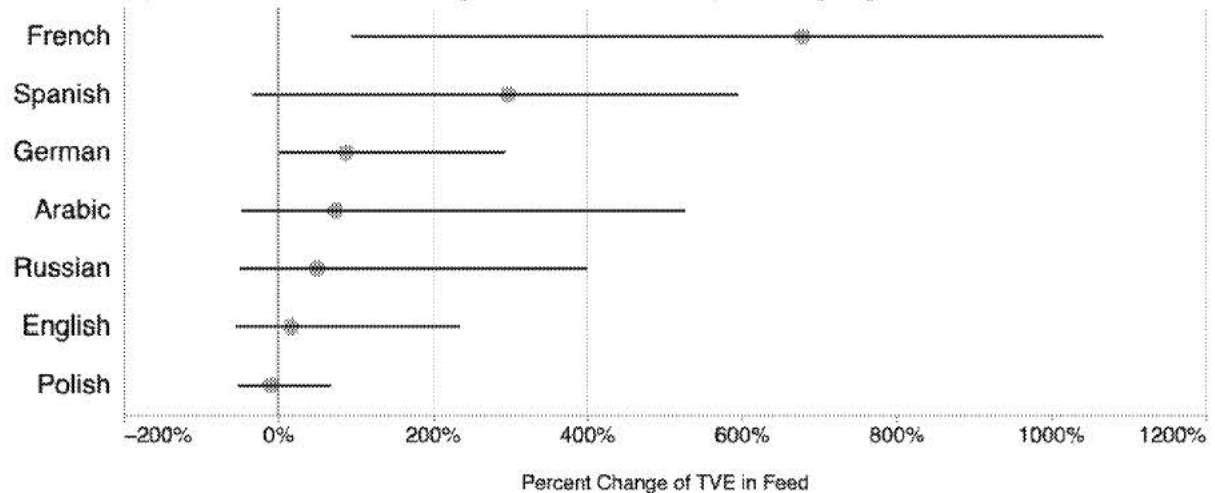


Figure A.6.11.7

This chart shows the percentage change for the amount of Bad (TVE) Content per language from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

Italian is missing from the above chart because there did not exist any Bad Content in feeds for the first search stage within the sample.

6.11.8 Amplification per Platform

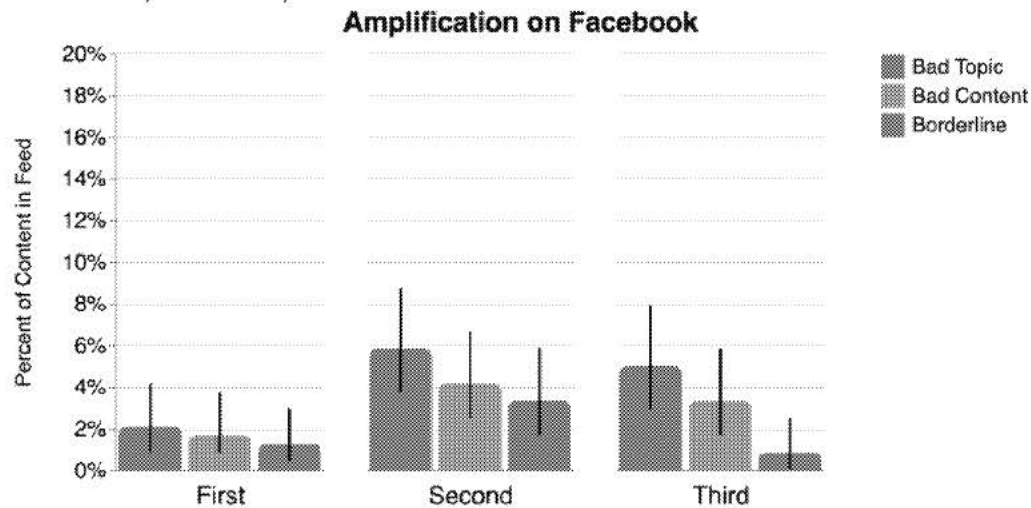


Figure A.6.11.8a

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Facebook for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification on Instagram

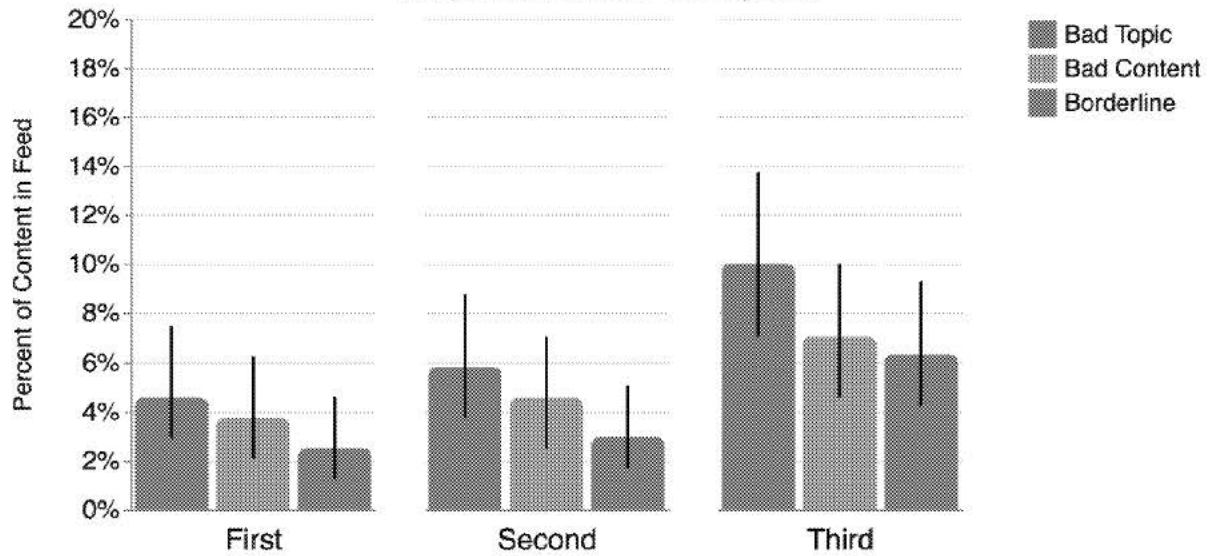


Figure A.6.11.8b

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Instagram for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification on TikTok

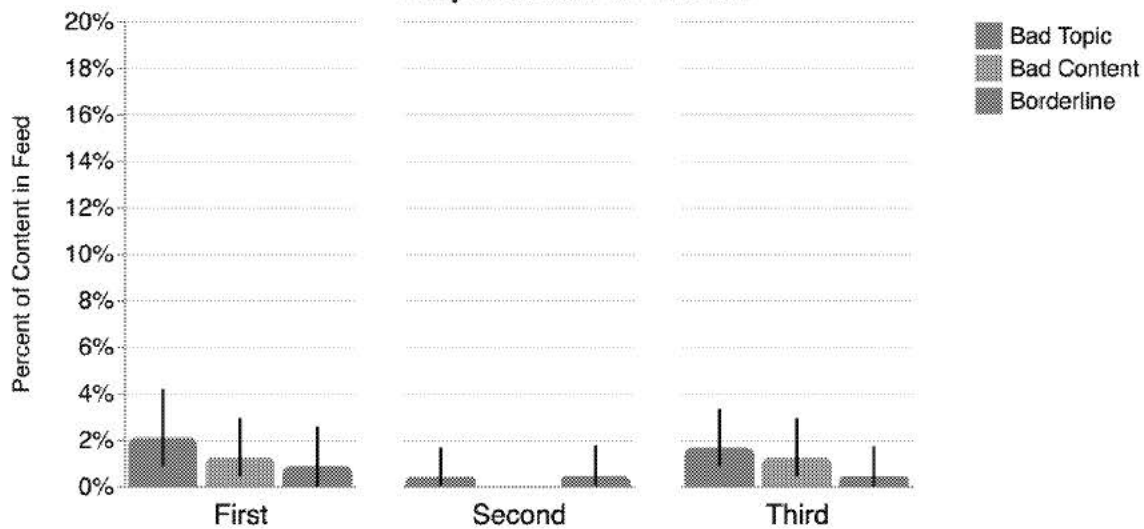


Figure A.6.11.8c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on TikTok for each stage. The black lines in the bars represent 90% confidence intervals.

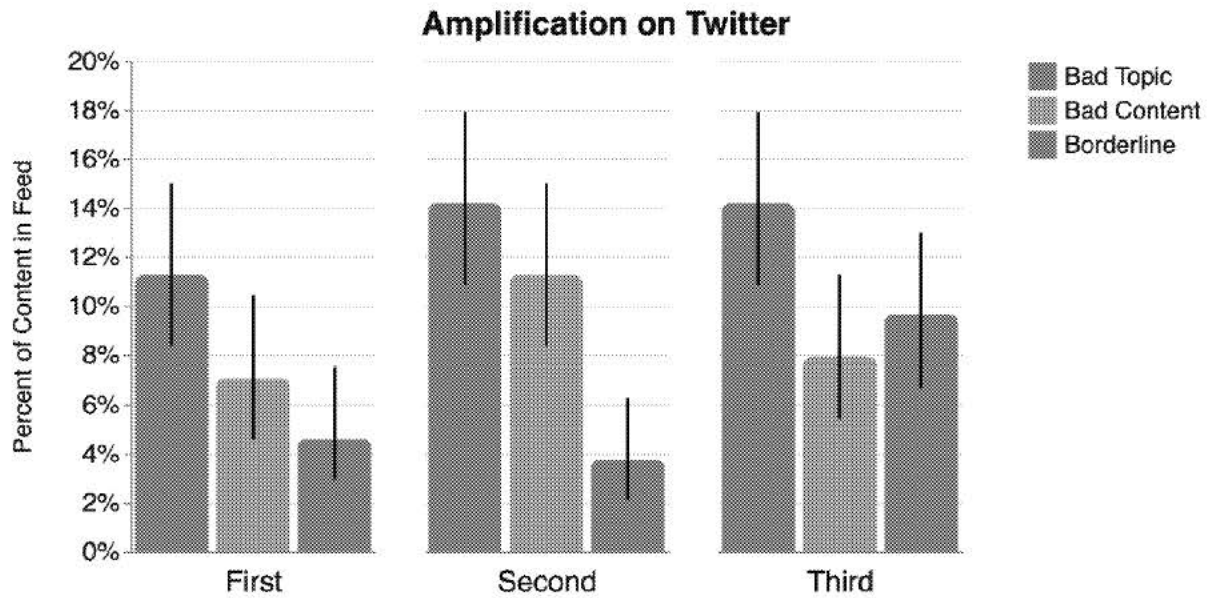


Figure A.6.11.8d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Twitter for each stage. The black lines in the bars represent 90% confidence intervals.

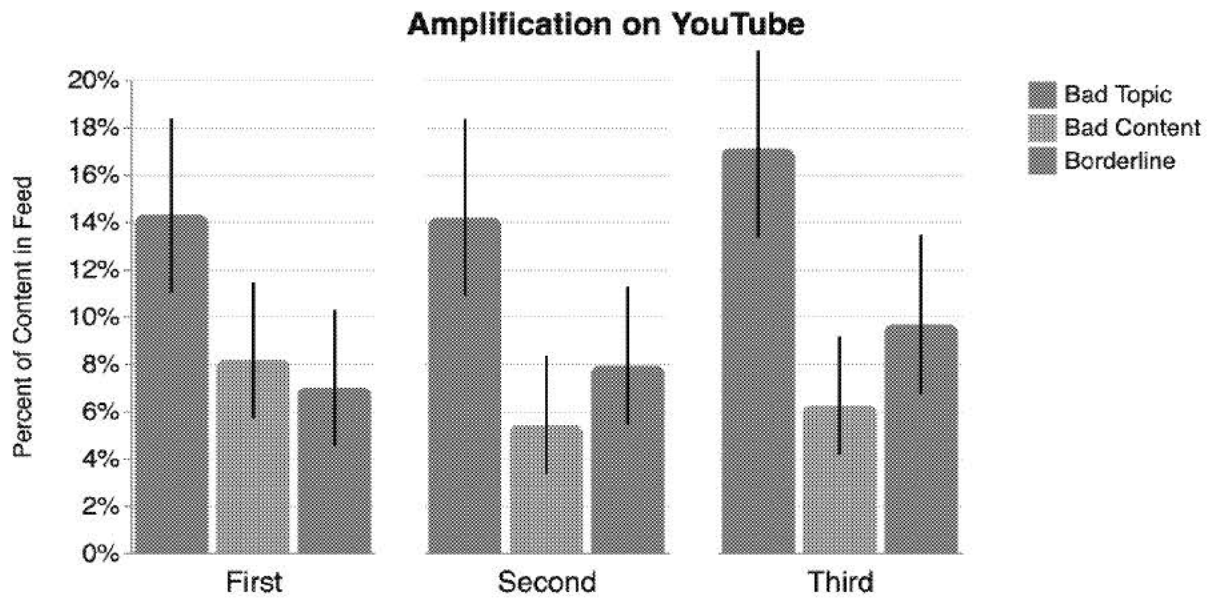


Figure A.6.11.8e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on YouTube for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.9 Amplification per Language

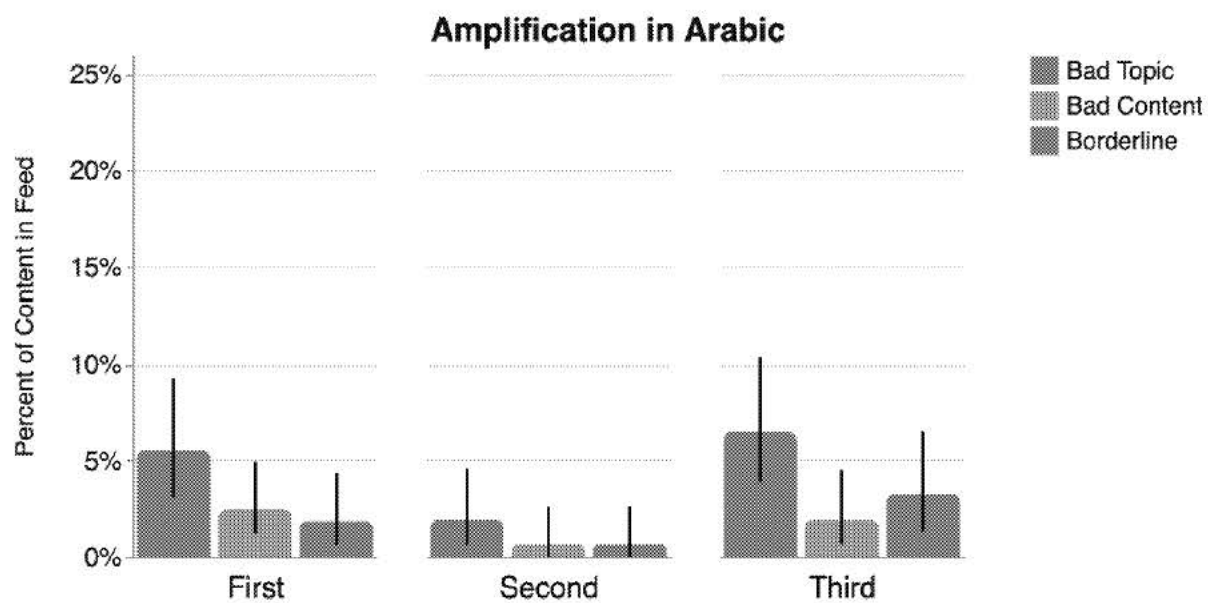


Figure A.6.11.9a

This graph illustrates the percentage of recommended content that is “Bad Content”, “Bad Topic”, and “Borderline” in Arabic for each stage. The black lines in the bars represent 90% confidence intervals.

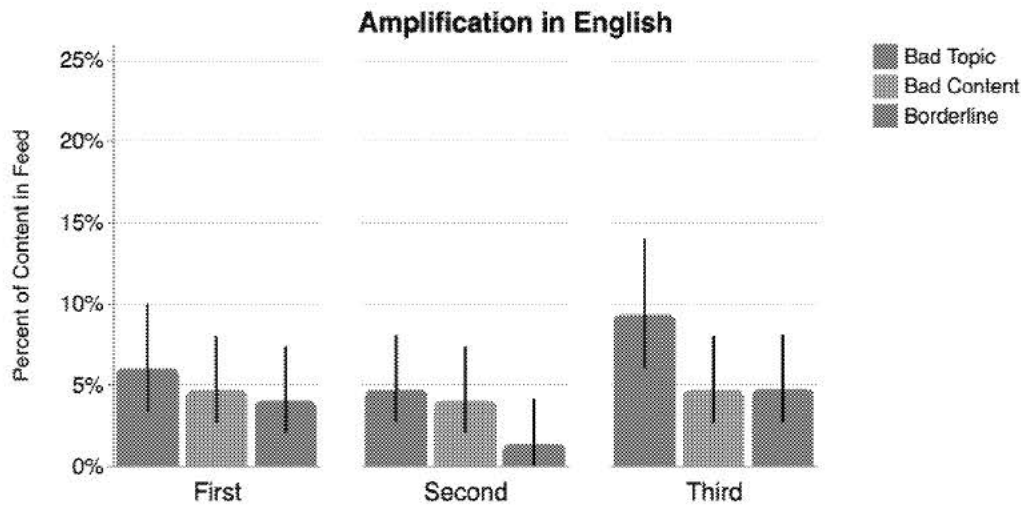


Figure A.6.11.9b

This graph illustrates the percentage of recommended content that is “Bad Content”, “Bad Topic”, and “Borderline” in English for each stage. The black lines in the bars represent 90% confidence intervals.

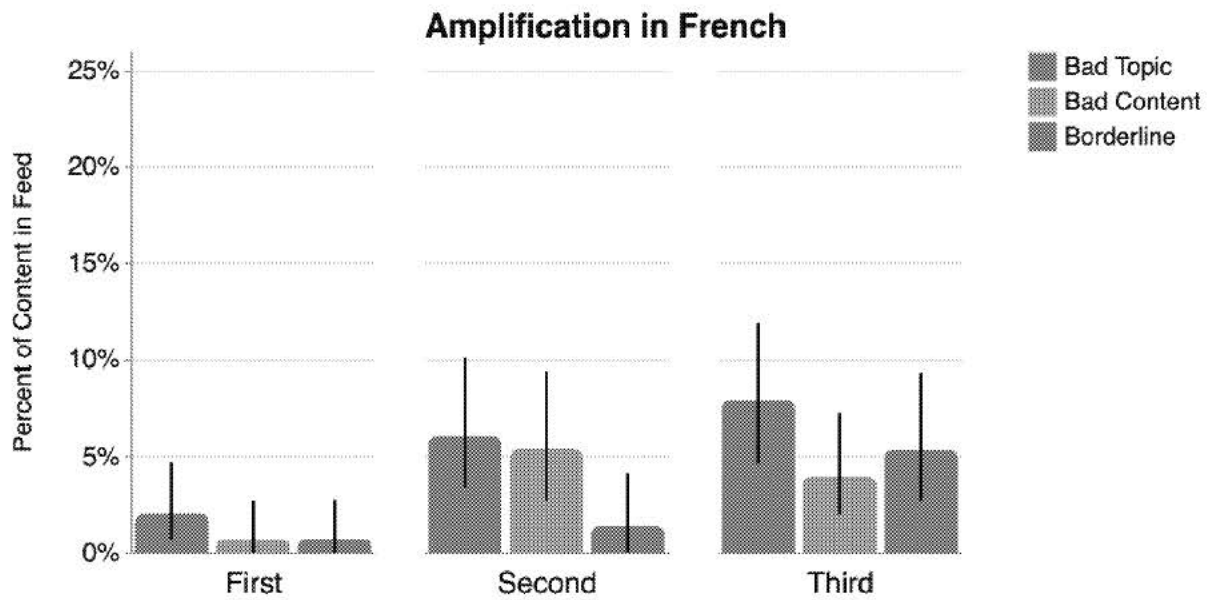


Figure A.6.11.9c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in French for each stage. The black lines in the bars represent 90% confidence intervals.

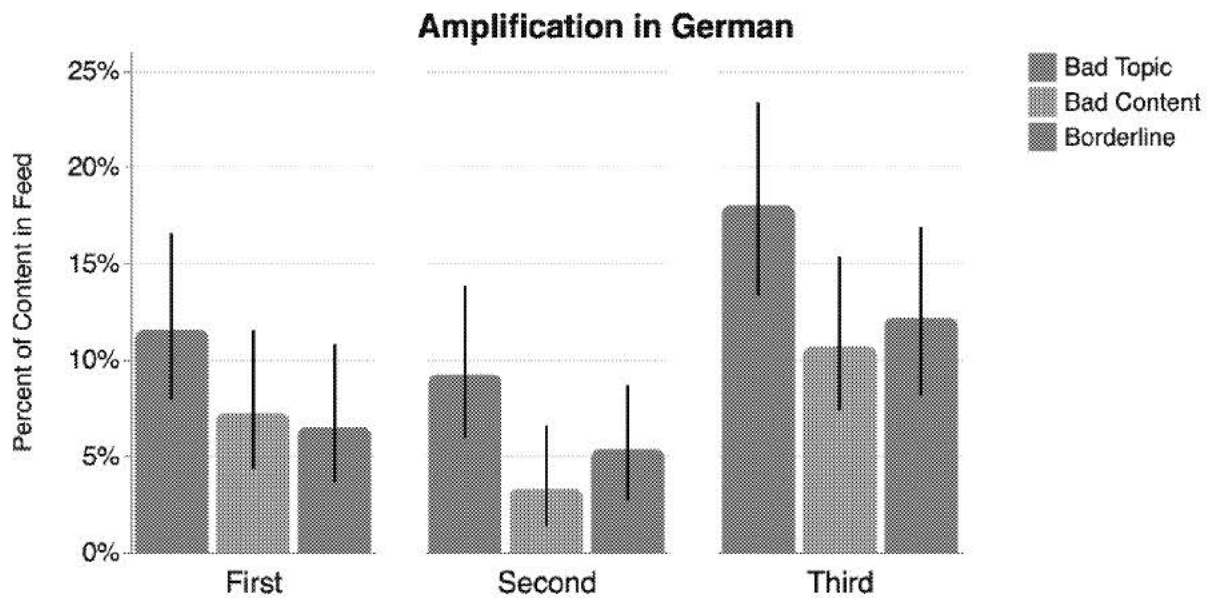


Figure A.6.11.9d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in German for each stage. The black lines in the bars represent 90% confidence intervals.

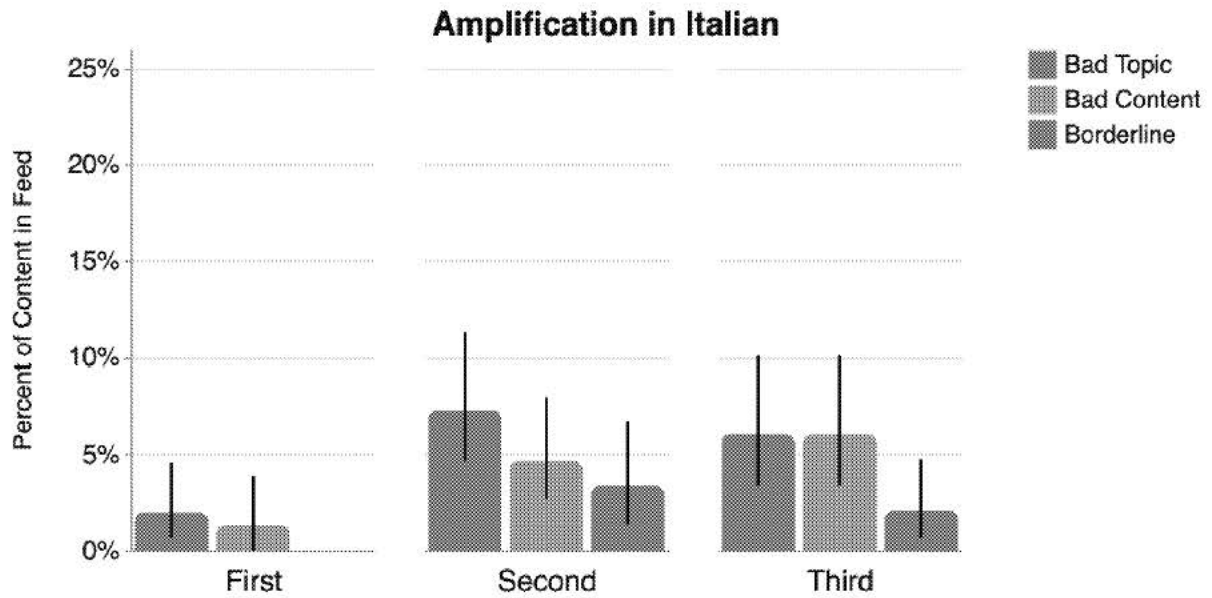


Figure A.6.11.9e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Italian for each stage. The black lines in the bars represent 90% confidence intervals.

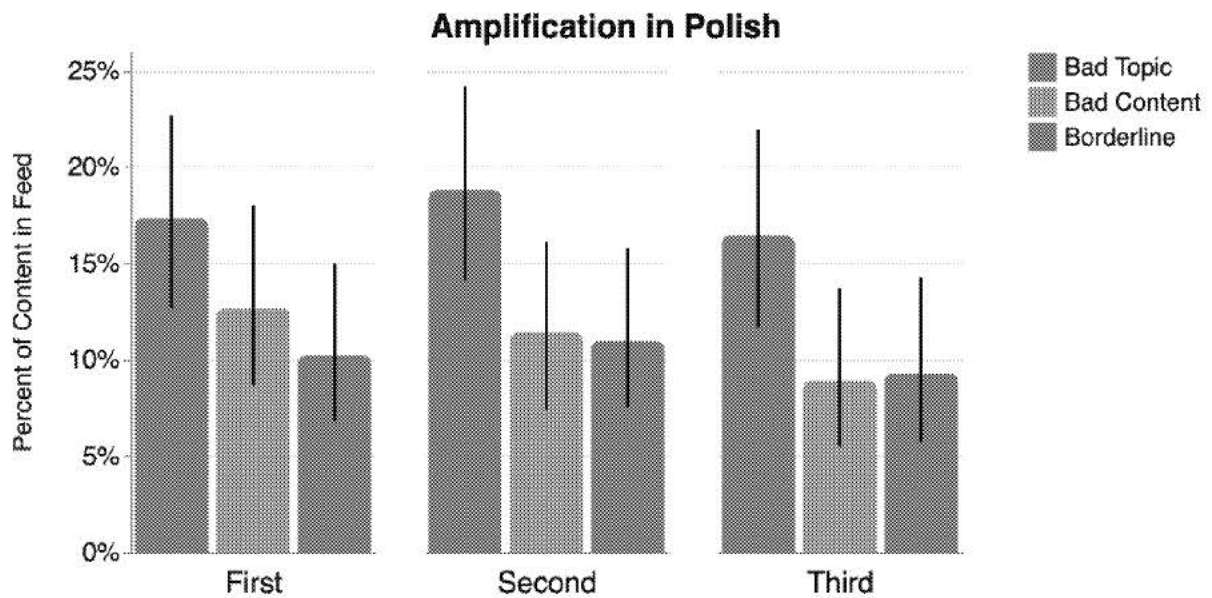


Figure A.6.11.9f

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Polish for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification in Russian

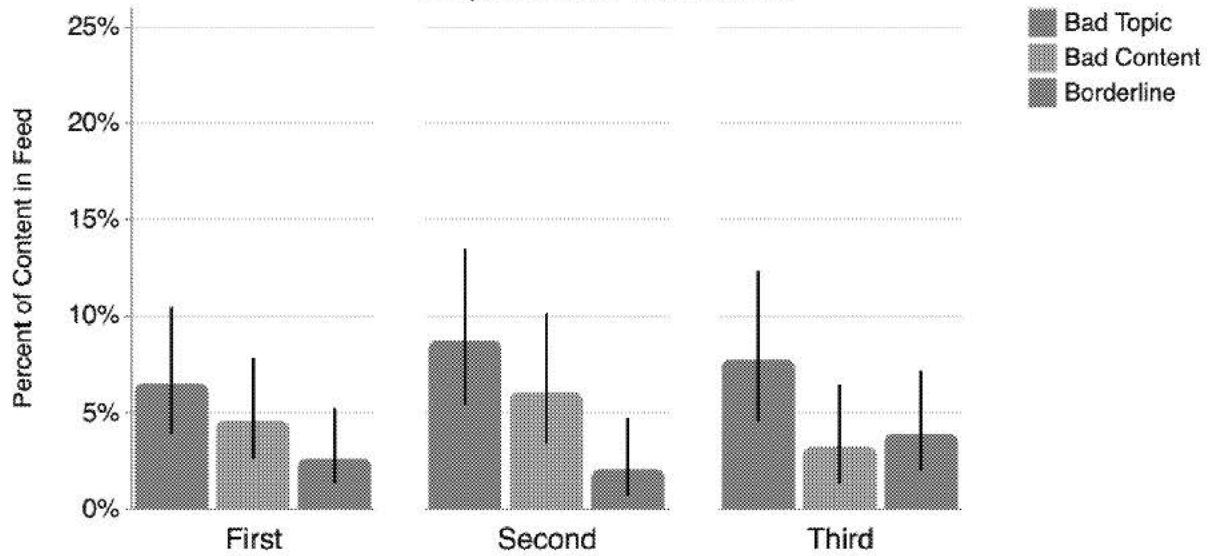


Figure A.6.11.9g

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Russian for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification in Spanish

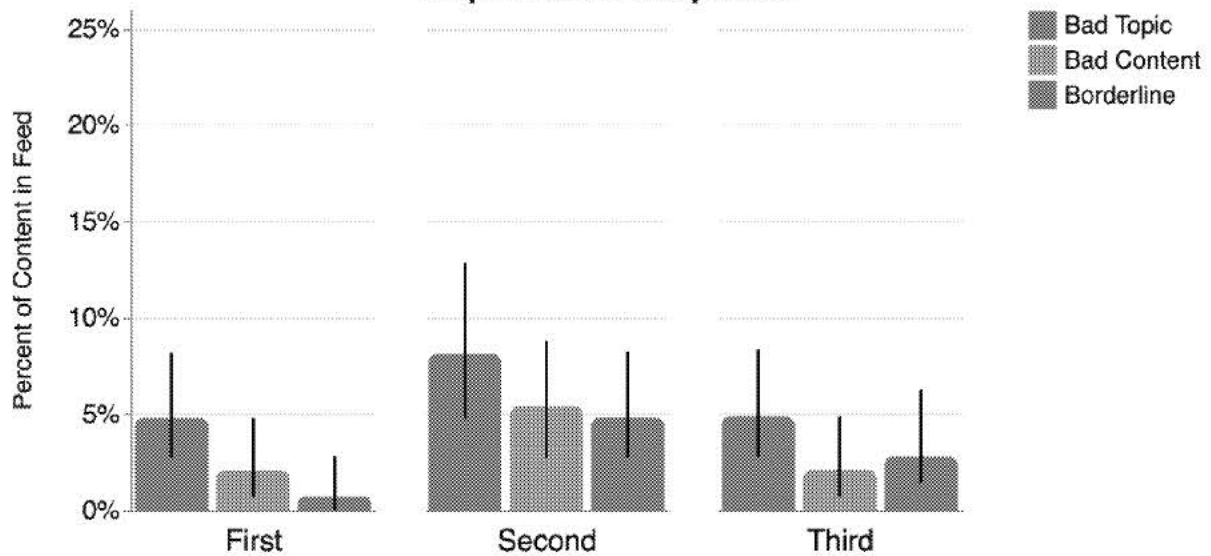


Figure A.6.11.9h

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Spanish for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.10 Amplification per TVE Type

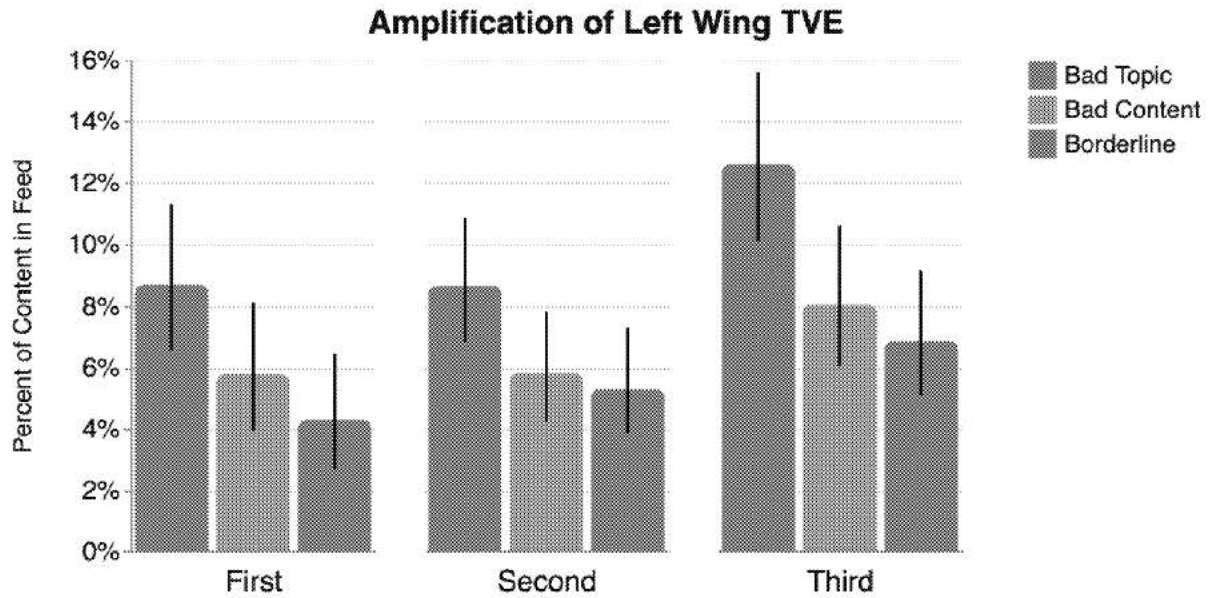


Figure A.6.11.10a

This graph illustrates the percentage of recommended Left-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

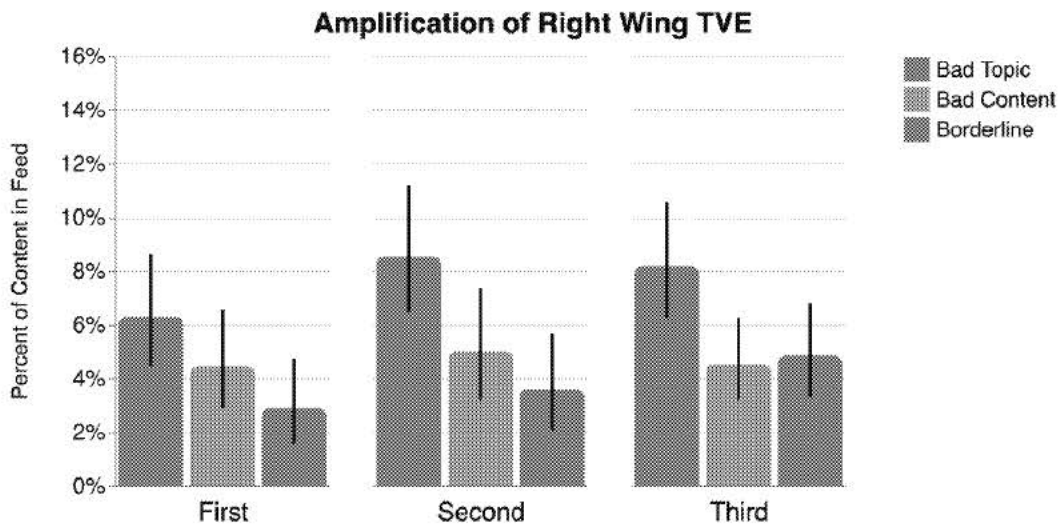


Figure A.6.11.10b

This graph illustrates the percentage of recommended National/Right-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

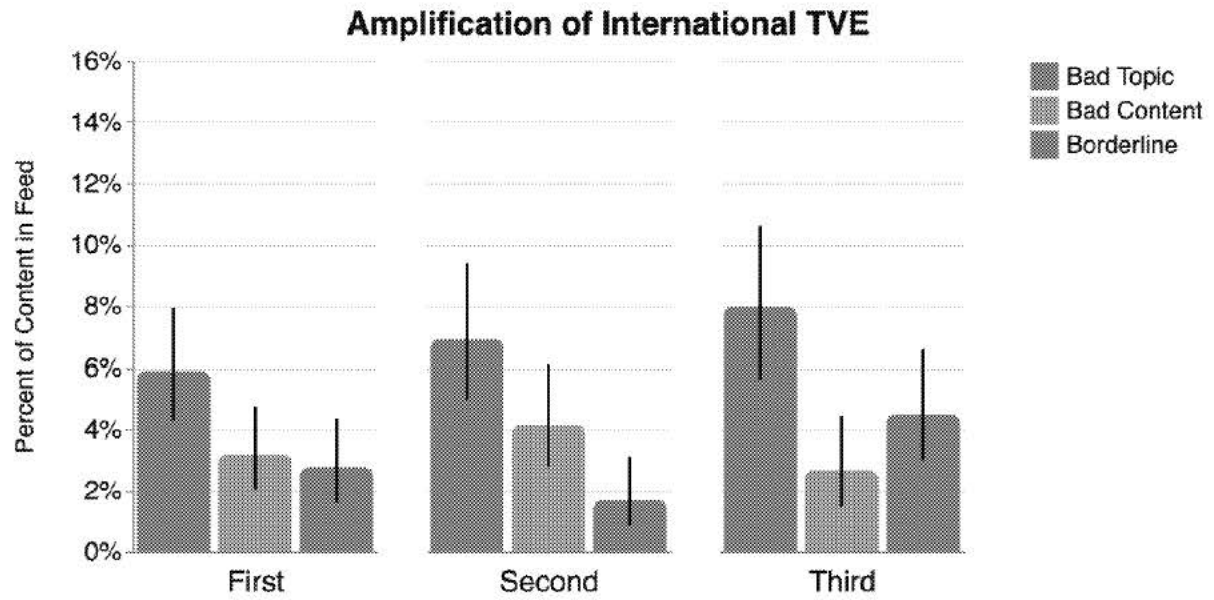


Figure A.6.11.10c

This graph illustrates the percentage of recommended international TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.11 Amplification P-values

Amplification Per Platform

Platform	Content Type	First/Second	First/Third	Second/Third
['Facebook']	Bad Topic	0.0352	0.0842	0.6875
['Facebook']	Bad Content	0.1041	0.2431	0.6317
['Facebook']	Borderline	0.1278	0.6538	0.0553
['Instagram']	Bad Topic	0.5387	0.0225	0.0913
['Instagram']	Bad Content	0.6486	0.1071	0.2435
['Instagram']	Borderline	0.7791	0.0447	0.0811
['TikTok']	Bad Topic	0.1007	0.7371	0.1782
['TikTok']	Bad Content	0.0826	1.0000	0.0826
['TikTok']	Borderline	0.5634	0.5634	1.0000
['Twitter']	Bad Topic	0.3384	0.3384	1.0000
['Twitter']	Bad Content	0.1142	0.7296	0.2156
['Twitter']	Borderline	0.6486	0.0328	0.0104
['YouTube']	Bad Topic	0.9701	0.3978	0.3799
['YouTube']	Bad Content	0.2306	0.4166	0.6976
['YouTube']	Borderline	0.6820	0.2908	0.5192

P-values on Amplification Average Percent by Content Type in Feed per Platform over time

Amplification per Language

108

Language	Content Type	First/Second	First/Third	Second/Third
['Arabic']	Bad Topic	0.0963	0.7269	0.0488
['Arabic']	Bad Content	0.1988	0.7540	0.3189
['Arabic']	Borderline	0.3441	0.4317	0.1014

Amplification per TVE Type

tve_type	Content Type	First/Second	First/Third	Second/Third
['International']	Bad Topic	0.5541	0.2520	0.5937
['International']	Bad Content	0.4608	0.6750	0.2793
['International']	Borderline	0.3139	0.1953	0.0317
['Left Wing']	Bad Topic	0.9888	0.0758	0.0538
['Left Wing']	Bad Content	0.9755	0.2102	0.1874
['Left Wing']	Borderline	0.4823	0.1127	0.3203
['National / Right']	Bad Topic	0.2485	0.2901	0.8635
['National / Right']	Bad Content	0.7282	0.9579	0.7547
['National / Right']	Borderline	0.6203	0.1643	0.3995

Table A6.11.11c

P-values on Amplification Average Percent by Content Type in Feed per TVE Type over time

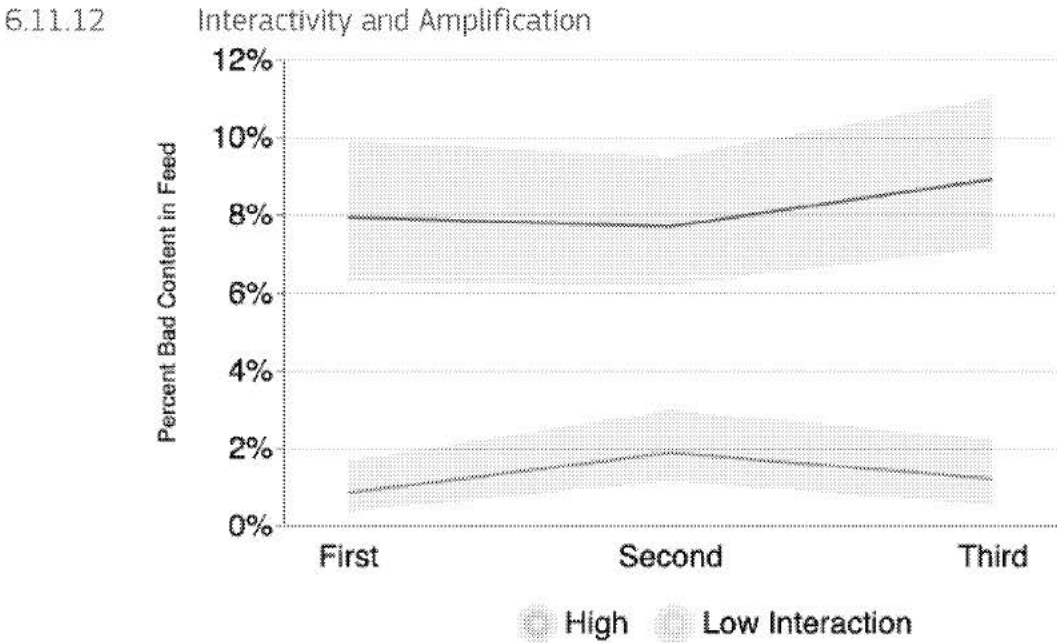


Figure A.6.11.12

This chart shows the average percentage of Bad Content for High and Low interaction over time. The shaded area represents the 90% confidence interval.

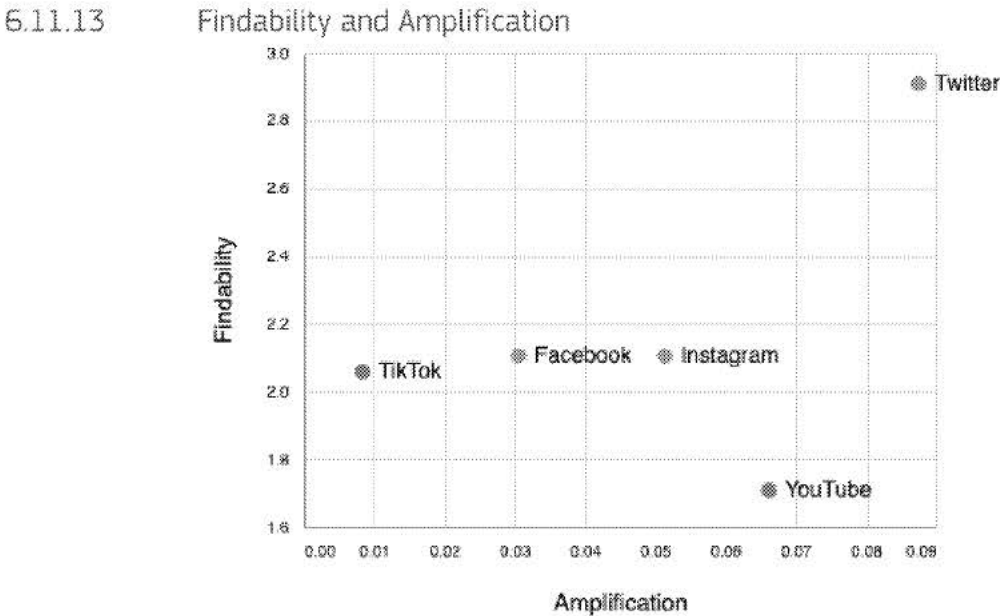


Figure A.6.11.13

This chart shows the comparison of the Findability score and the average percentage of Bad Content in the feed (here called Amplification). TikTok and YouTube are highlighted in red to contrast tech investment in feed vs search.

6.11.14 Engagement Ratio and Amplification

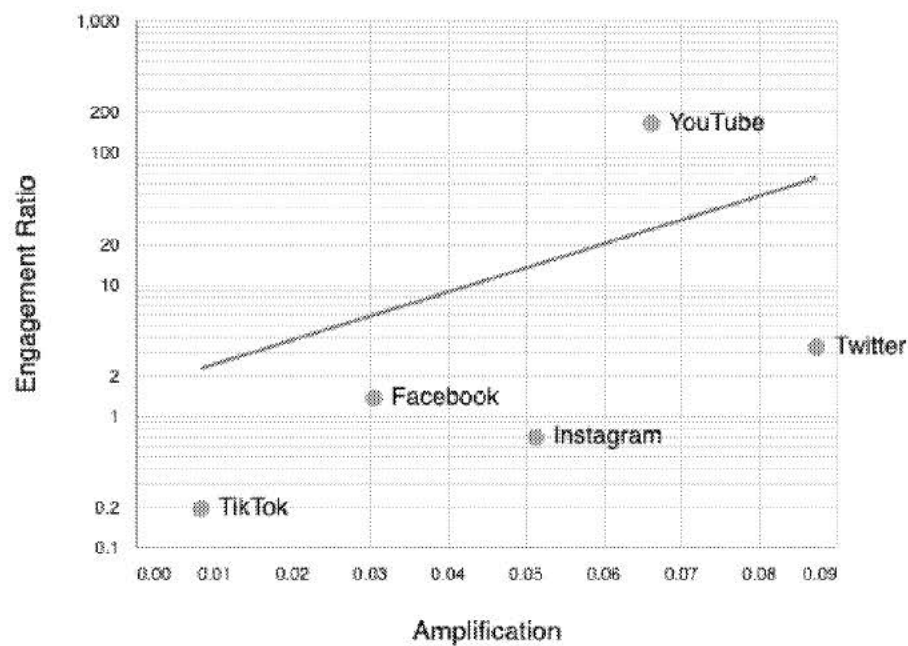


Figure A.6 11.14

This chart shows the positive correlation of the engagement and amplification. To normalise the engagement, an engagement ratio was used which is the median engagement (sum of the likes, shares, comments, and followers) of TVE content divided by the median engagement of non-TVE content. Amplification is the average proportion of the feed that contains TVE.

(note the log scale for the engagement ratio)

6.11.15 Removal Rate and Amplification

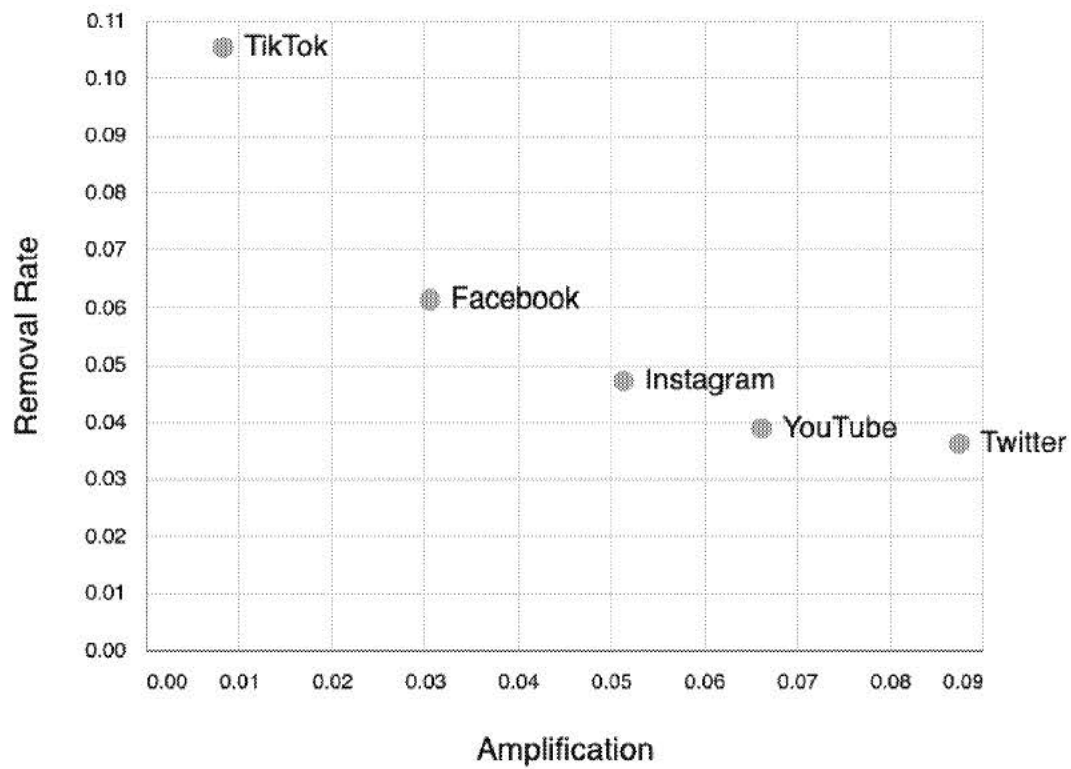


Figure A.6.11.15

This chart shows the inverse correlation of the removal rate and amplification.

6.12 Borderline Charts

6.12.1 Removal Rates per Platform

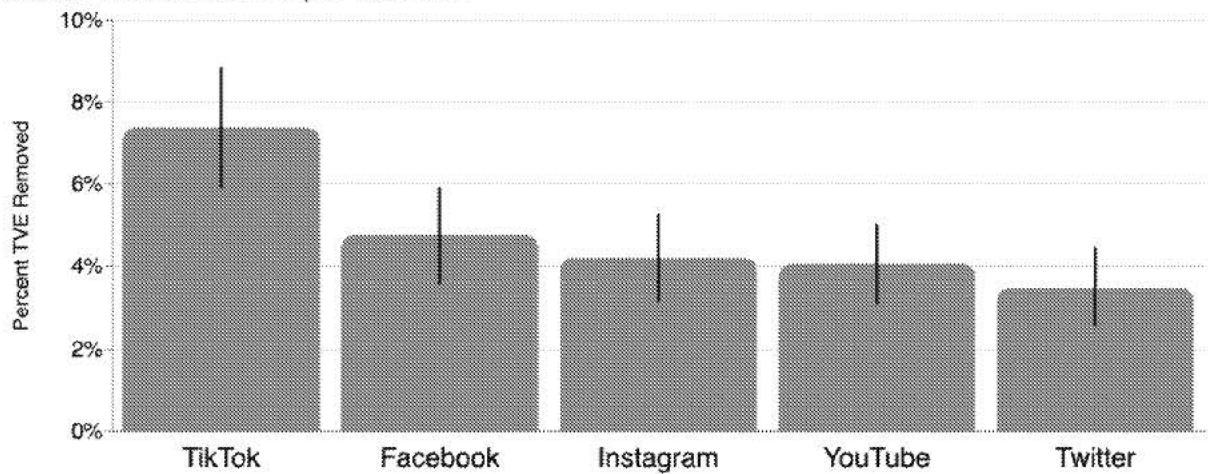


Figure A.6.12.1
This chart shows the percentage of Borderline TVE content that was removed by each platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.6397	0.9144	0.0505	0.6846
Twitter	0.4009	0.6201	0.0237	
TikTok	0.1604	0.0774		
Instagram	0.7283			

Table A.6.12.1
P-values on Removal Rates per Platform (Borderline)
 $\alpha = 0.01$ (Bonferroni correction from 0.05)

6.12.2 Removal Rate per Language

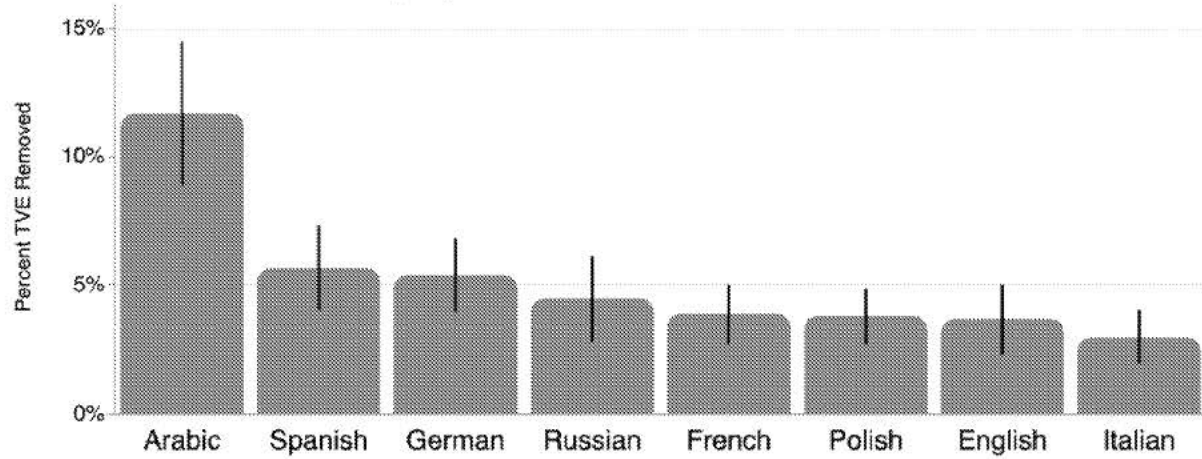


Figure A.6.12.2

This chart shows the percentage of Borderline TVE content that was removed for each language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0507	0.3527	0.3581	0.8964	0.1516	0.3207	0.6160
Russian	0.0226	0.7029	0.7601	0.6869	0.4226	0.7209	
Polish	0.0014	0.9431	0.9576	0.3679	0.5937		
Italian	0.0005	0.6862	0.5672	0.1741			
German	0.0281	0.3993	0.4095				
French	0.0023	0.9076					
English	0.0052						

Table A.6.12.2

P-values on Removal Rates per Language (Borderline)

$\alpha = 0.00625$ (Bonferroni correction from 0.05)

6.12.3 Removal Rate per TVE Type

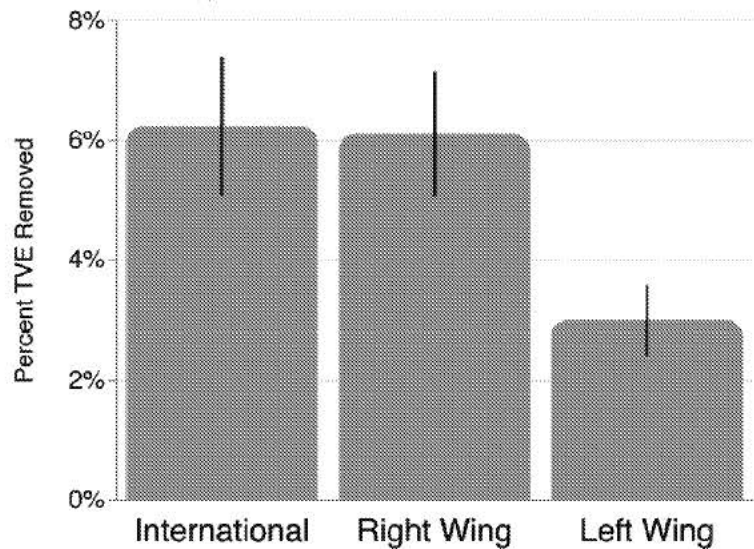


Figure A.6.12.3

This chart shows the percentage of Borderline TVE content that was removed for each TVE Type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.9357	0.0051
Left Wing	0.0056	

Table A.6.12.3

P-values on Removal Rates per TVE Type (Borderline)

alpha = 0.017 (Bonferroni correction from 0.05)

6.12.4 Removal Time per Platform

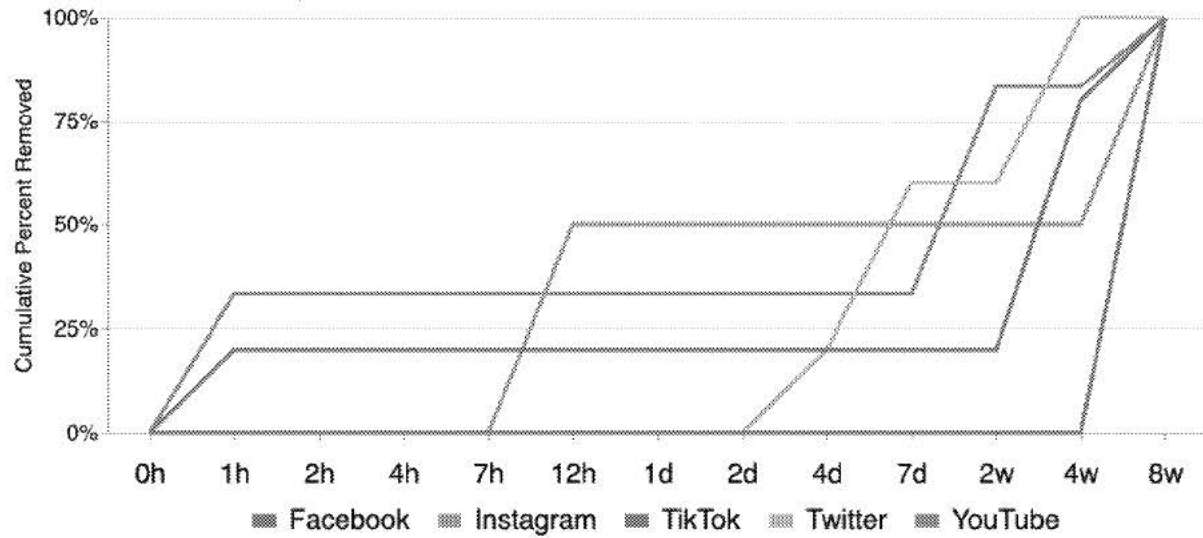


Figure A.6.12.4

This chart shows how long it took for each platform to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.5 Removal Time per Language

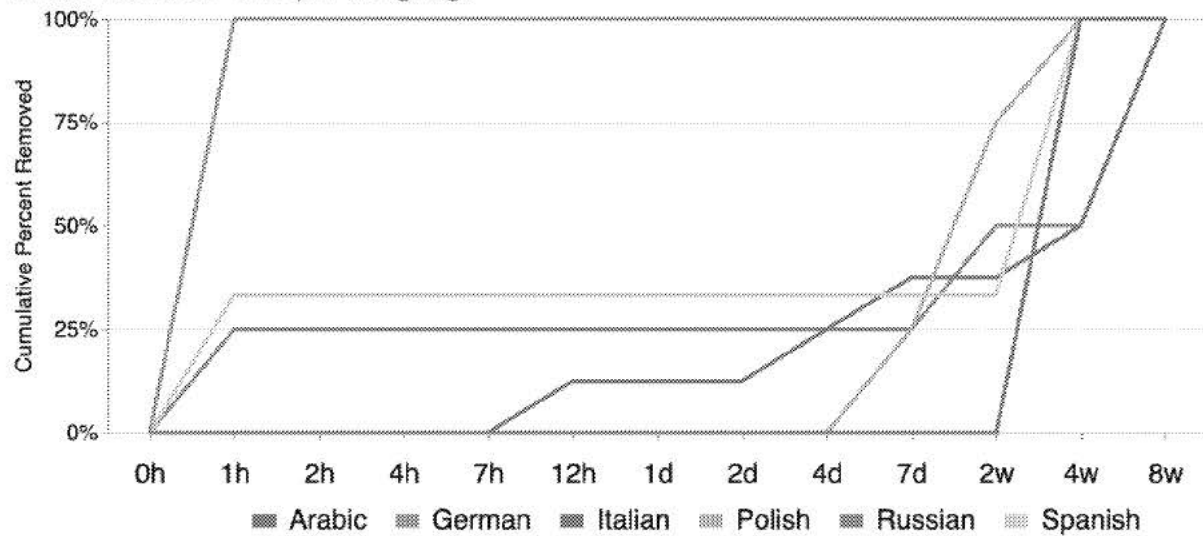


Figure A.6.12.5

This chart shows how long it took for each language to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.6 Removal Time per TVE Type

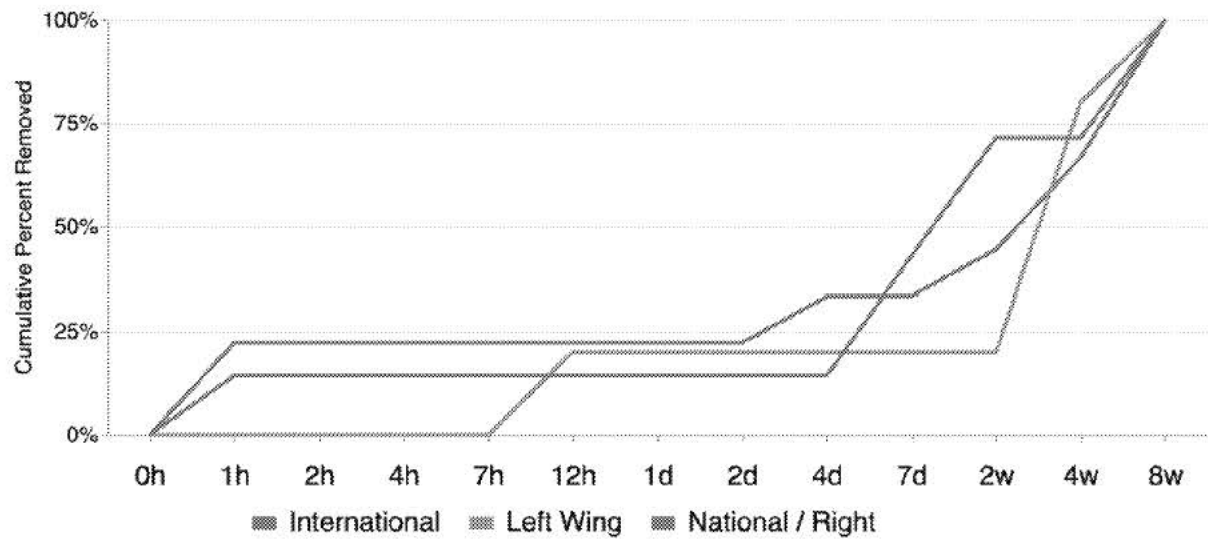


Figure A.6.12.6

This chart shows how long it took for each TVE Type to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.13 Task 3: List of References

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410-428.

Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78, 413-418.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*, 10(21), 7748.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945.

Grimmelmann, J. (2015) The virtues of moderation. *Yale Journal of Law & Technology* 17: 42.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3), 261-273.

Mohamed, M. H., Khafagy, M. H., & Ibrahim, M. H. (February 2019). Recommender systems challenges and solutions survey. In 2019 international conference on innovative trends in computer engineering (ITCE) (pp. 149-155). IEEE.

Murthy, D. (2021). Evaluating platform accountability: terrorist content on YouTube. *American behavioral scientist*, 65(6), 800-824.

Suhaim, A. B., & Berri, J. (2021). Context-aware recommender systems for social networks: review, challenges and opportunities. *IEEE Access*, 9, 57440-57463.

Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7, 439-457.

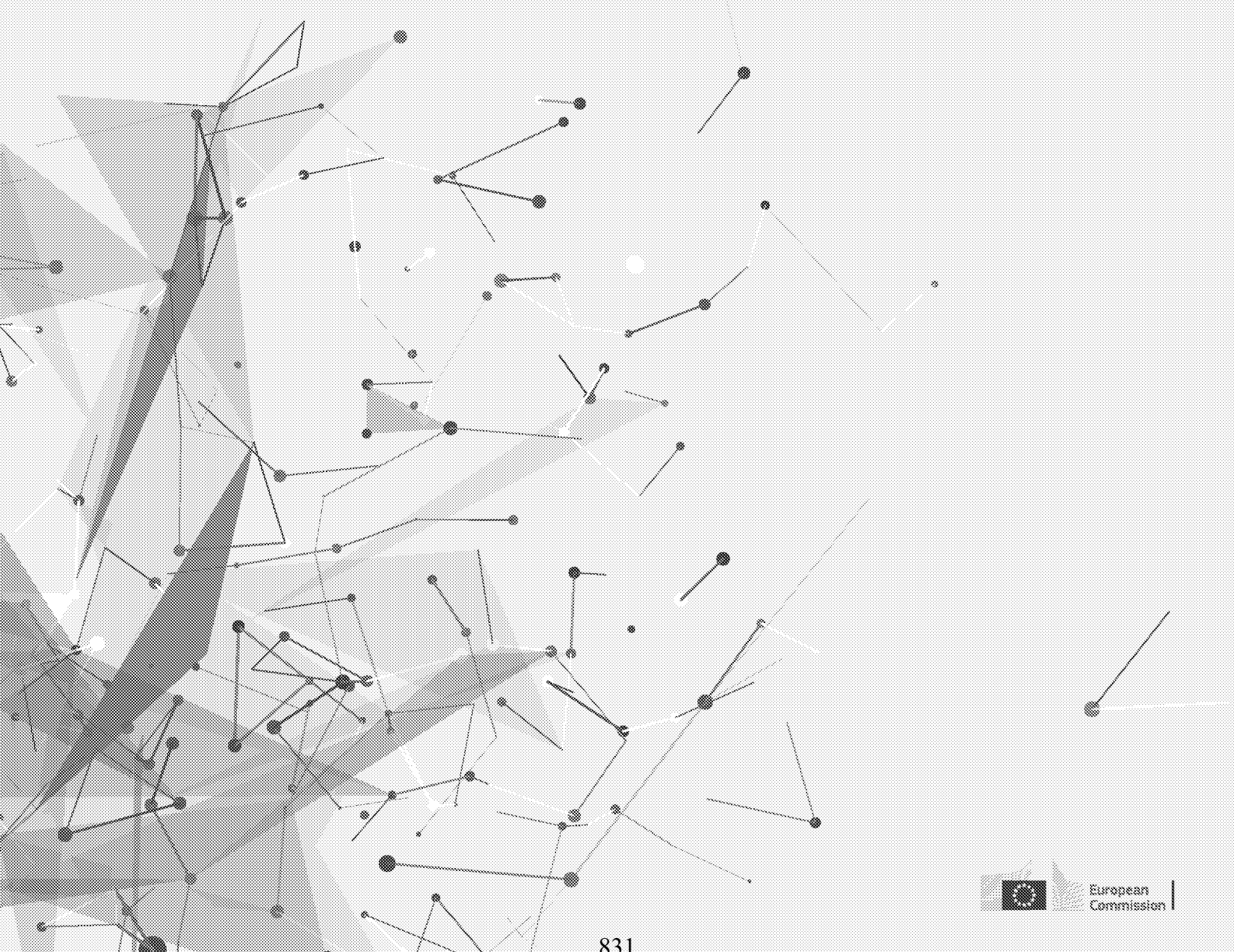
END OF THE DOCUMENT

Exhibit 38

EU Internet Forum

THE HANDBOOK OF BORDERLINE CONTENT IN RELATION TO VIOLENT EXTREMISM

Definitions, examples and preventive measures to identify harmful content leading to radicalisation and violent extremism



This **EUIF handbook** was drafted in cooperation with the **Global Internet Forum to Counter Terrorism (GIFCT)** and includes contributions from:

EU Services, agencies and contractors: DG HOME, DG JUST, DG CNECT, the Fundamental Rights Agency (FRA), EU External Action Service (EEAS), EU IRU at Europol, Radicalisation Awareness Network (RAN).

EU Member States: Germany, Italy, France, the Netherlands, Romania, Czech Republic, Spain, Slovakia.

Civil Society Organisations, think tanks and researchers: Tech Against Terrorism, Christchurch Call Advisory Network (CCAN), Institute for Strategic Dialogue (ISD), Center for Countering Digital Hate (CCDH), GNET, Counter Extremism Project (CEP), Alexander Ritzmann, Lisa Kaati.

Disclaimers

1. *Reproduction of the hateful terms and items listed in this handbook should not be interpreted as any kind of endorsement of the views and sentiments expressed within. They appear only to replicate original content as seen online. The intention behind their inclusion is to inform EUIF members of the threats such content can pose to users.*

Similarly all disinformation narratives and related items which appear are included for the sole purpose of alerting readers to their harmful nature.

2. *This handbook represents a collection of operational information, case studies, research and analysis provided by members of the EUIF with the objective of presenting various perspectives on borderline content in relation to violent extremism. It is aimed for use by tech companies when moderating this harmful content. The analysis and recommendations contained within **do not represent** the views of the European Commission.*

Index

Introduction	5
What is Borderline Content in relation to Terrorist and Violent Extremist Content?.....	7
Definitions and approaches of borderline content in relation to violent extremism and radicalisation.....	7
Main takeaways of EUIF discussions.....	7
Post organisational violent extremism and use of borderline content	9
Terrorgram.....	10
GIFCT: defining borderline content in relation to violent extremism and processes of radicalisation	11
Borderline Content Sub-Theme Policies and Use in TVE and Non-TVE Cases	13
EUIF review of Analysis and definitions of Dehumanisation.....	14
Categories of borderline content in relation to radicalisation and violent extremism.....	17
Racist, anti-migrants and antisemitic content.....	17
Covid-19 crisis-related borderline content.....	20
Pro-Kremlin borderline content.....	24
The use of irony and memes.....	29
Borderline content linked to the Bratislava attack.....	30
Anti-LGBTIQ and anti-feminist content	33
Policy perspective: what are the current measures in place to prevent the spread of borderline content	38
Legislative and voluntary approaches in the EU.....	38
The voluntary/cooperative approach.....	38
The legislative approach.....	39
Applying definitions of borderline content in relation to violent extremism into policy guidelines	42
Suggested guidelines	46
Definitions, taxonomies and content moderation.....	46
Risk assessments	50
Mitigation measures: the role of recommender systems	51
Conclusion	56
Box: GIFCT Review and Analysis of Tech Company Efforts to Counter Borderline Content	
Annex I	58
EUIF Handbook's Glossary	
Annex II.....	63
Box: GIFCT Review and Analysis of Tech Company Efforts to Counter Borderline Content	

Introduction

The EU Internet Forum's key objective is to work hand-in-hand with the industry, EU Member States, Europol and civil society to develop instruments to jointly prevent the dissemination of harmful content leading to offline violence.

The term 'borderline content' has become increasingly prominent 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)**, the **Global Internet Forum to Counter Terrorism (GIFCT)**, and the **Christchurch Call to Action**. In these fora, governments, technology companies, and expert stakeholders have implemented joint processes and deliverables to understand and develop concrete measures to counter terrorism and violent extremism online.

A common understanding on 'borderline content' in relation to processes of radicalisation and violent extremist content is difficult to pinpoint as it is difficult to agree on general differences with respect to the treatment of illegal and legal (but potentially harmful) content, such as **disinformation, conspiracy theories**, and other types of content that is not labelled as violent extremist but can nonetheless contribute to a process of radicalisation towards violence. As regards those forms of **hate speech** potentially leading to violent extremism that are illegal in EU Member States (for instance, content that incites racist and xenophobic violence and hatred), it is important to underline that they are not always identified as illegal by tech platforms, but rather identified as 'grey zone' content, due to the use of coded language or to a lack of clarity at their author/amplifier's intentions.

Against this backdrop, this handbook aims at raising awareness about the need for making clearer links between hate speech and terrorist and violent extremist incidents or activities. It has the sole purpose of providing **non-legally binding guidance** on how to better **understand and respond to borderline content that may lead to radicalisation and violent extremism**. Its objective is to better understand the **links between this content and violent extremism**. It does not provide guidance on the legality of such content under EU or national laws, for instance to counter racist, xenophobic, homophobic or misogynist violence and hatred.

The European Union and its Member States are bound by the human rights obligations enshrined in the Charter of Fundamental Rights of the European Union (hereafter 'the Charter') and the European Convention on Human Rights (ECHR)¹ freedom of expression², and information³. In that context **any measures at EU level addressing borderline content in relation to radicalisation and violent extremism need to be based on the full respect of fundamental and human rights**.

In that respect, regarding 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, as enshrined in the Charter.

The handbook is the result of multi-stakeholder exchanges within the EUIF, in which all parts 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. All the analysis and information contained in this handbook was provided by EU services and agencies, GIFCT, EU Member States, civil society organisations, as well as key external stakeholders and researchers with a strong expertise in this field. The first part of the handbook is meant to provide existing **definitions and taxonomies**

1 As per Article 11 of the Charter of Fundamental Rights on Freedom of expression and information [LINK](#)

2 Article 11 of the Charter of Fundamental Rights of the European Union [LINK](#)

3 Article 6 of the Charter of Fundamental Rights of the European Union [LINK](#)

of borderline content, with the important support of the Global Internet Forum to Counter Terrorism (see Annex II), and includes **concrete examples** of content that EU Member States, researchers and civil society organisations have identified as potentially leading to violent extremism. The second part aims at providing information on **EU legislative actions** that may have an impact on the spread of borderline content related to TVEC and **tech platforms'** guidelines and internal policies. In the last section, the handbook gathers **recommendations** provided by Member States, EU Services and civil society organisations on how to jointly limit the spread of borderline content. While some definitional framing and conclusions from GIFCT are within the body of the handbook, GIFCT provides further review and analysis of member companies' capacities and efforts to counter borderline content provided in the Annex.

To facilitate the use of this handbook by tech companies in their content moderation efforts, a **glossary** is included with **brief descriptions, images and keywords** associated to each category of borderline content identified in the handbook.

As the EUIF observed how borderline content is used to legitimate and normalise financing activities, this handbook also provides some guidance for **tech companies on how to identify the misuse of their platforms for financial gains and provide recommendations on how to prevent the sale of merchandise promoting dangerously hateful and extremist ideologies.**

With the precious **support of GIFCT and other members** of the Forum, the EUIF will continue to explore ways and provide instruments to prevent the spread of illegal and harmful content and behaviours. This handbook will be updated on a yearly basis.

Introductory Framing by the Global Internet Forum to Counter Terrorism on its Contribution⁴

GIFCT is a non-profit organisation with the mission of preventing terrorists and violent extremists from exploiting digital platforms. To further this aim in relation to 'borderline content', rather than calling for a unified definition or criteria of actions against borderline content, GIFCT 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.

Ultimately, 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.

4 GIFCT contribution to the EU Internet Forum Handbook on Borderline Content. Any feedback, questions, or comments can be emailed to erin@gifct.org and micalie@gifct.org

What is Borderline Content in relation to Terrorist and Violent Extremist Content?

Definitions and approaches of borderline content in relation to violent extremism and radicalisation

Main takeaways of EUIF discussions

What emerged from the exchanges and activities held by the EU Internet Forum on borderline content, such as the workshop on algorithmic amplification and borderline content held on 29 September 2022, is that the type of harmful content that should be considered as potentially leading towards radicalisation consists in a **combination of disinformation/conspiracy theories and forms of hateful content**, which is not automatically identified by online platforms as illegal or as potentially conducive to violent extremist or terrorist acts.

In the exchanges held during the EUIF workshop in September 2022, EU Member States and members of civil society outlined a number of categories and examples of harmful but lawful content used by violent extremists and terrorists to spread their propaganda and radicalise users.

As regards the specific nature of the **borderline content** leading to violent extremism, most experts and policy makers point at the following clusters: anti-establishment/anti-institutions, antisemitic⁵, anti-LGBTIQ, misogynistic, anti-migrants, racist, anti-COVID measures.

A number of EU Member States provided guidelines and examples of either borderline content or illegal hate speech that are not always linked to violent extremism and therefore can be difficult to identify. The content was classified within the following categories:

- **Grey zone antisemitic** imagery and Holocaust denial distortion
- **Grey-zone anti-migrants** and Islamophobic content
- **Anti-LGBTIQ and misogynist** content
- Visual propaganda regarding **non-proscribed extremist groups**, usually right-wing leaning ones
- **Misleading and deceptive content** related to the **Ukrainian conflict**, including revisionist rhetoric and anti-European/anti-West and anti-immigration narratives
- **Anti-government/system** content meant to incite violence

⁵ Antisemitic content can be illegal, when it falls under the categorisation of 'using degrading, defamatory words/expressions to name certain social groups/individuals belonging or perceived to belong to such groups', which also implies the glorification and denial of certain events important to the group. However, as explained in the introduction, such content is not always easy to detect and qualify, due to the use of tactics by antisemitic users to evade automatic detection and remain on the grey-zone between illegal content and content protected by freedom of speech. Even when it is not illegal, antisemitic content is still harmful. The Commission uses the non-legally binding working definition of antisemitism of the International Holocaust Remembrance Alliance (IHRA definition) as a practical guidance tool and a basis for its work to combat antisemitism.

EU Member States also indicated the following:

- **Violent right wing** extremists use and disseminate borderline content to promote their agenda, recruit new cadres and evade detection and content moderation efforts.
- Moreover, borderline content leading to violent extremism is very **diverse (images, videos, texts, gifs etc.)** and is always adapted to the social, political and cultural context of target users, often taking advantage of grievances and crisis situations as a means to reinforce their messages.
- Borderline content is not specific to one online space and it usually **spreads across platforms and spaces**.
- Borderline content also has a strong **transnational dimension**.

As regards malicious actors spreading borderline content, the Netherlands suggested they can be divided into 4 different categories:

- Political violent movements
- Violent movements against policy decisions
- Violent movements against the so-called 'elite'
- Copycats

The most impactful category is the third category (violent movements against the so-called 'elite'). However, within this category, it is important to make a distinction between public disorder and anti-institutional violent extremism, which causes long term concerns about a potential undermining of democracy and the escalation of violence.

It was also suggested by governments to make a **distinction** between **influencers and material perpetrators** and to look into the tactics used by intellectual online influencers active online, who lead the discussion (using persistent narratives against the elites and institutions) and create social media strategies to evade content moderation. When looking into the activities of influencers, one must not only look at how many followers they have, but also at the amount of views of the content itself. The so-called 'Material perpetrators' are the users who are publishing the content, while platforms are responsible for making it visible.

The use of borderline content is among the **categories of tactics used by violent extremists to evade moderation** that were identified by the Institute for Strategic Dialogue (ISD)⁶ in a paper funded by and produced for the Global Internet Forum to Counter Terrorism (GIFCT) website titled 'A Taxonomy for the Classification of Post-Organisational Violent Extremist & Terrorist Content':

- Use of confidential communication platforms
- Use of restricted chats, i.e. satellites, regional, etc.
- Use of sophisticated tools for account anonymization
- Use of fake news and borderline content to avoid to be pursued
- Use of borderline content to fund terrorist groups

6 J. Davey, M. Cornerford, J. Guhi, W. Baidet, C. Colliver, *A Taxonomy for the Classification of Post-Organisational Violent Extremist & Terrorist Content*, <https://www.isdglobal.org/wp-content/uploads/2022/01/A-taxonomy-for-the-classification-of-post-organisational-terrorist-content.pdf>

The EUIF is also dedicating efforts to further explore the use of borderline content to raise funds for violent extremist groups online. The nature of violent extremist and terrorist financing has changed in recent years, with Europol noting a clear **upsurge in different financing activities online**. These activities can take the form of crowdfunding and fundraising campaigns, the sale of merchandise as well as influencer activities, such as paid question and answer sessions or livestreams with income from advertisement. Merchandise often takes the form of clothing, articles, jewellery, bags, stickers and buttons as well as mugs. A key challenge in addressing this type of financing is that this content often does **not cross the threshold to be considered as illegal**. Many activities deliberately fall under the approach of ‘extreme normalisation’, where actors avoid explicitly illegal activities and language to circumvent the removal of the content or the disabling of access to it by the intermediary service. This also carries the danger that regular users buy items without understanding the context and the political and ideological significance of the item. Thereby unitedly promoting dangerously hateful and extremist ideologies.

Despite being below the **threshold of illegality**, these activities generate financial profits, account for legal defence, finance lifestyles of extremists, and directly **finance propaganda efforts leading to radicalisation and recruitment** and in extreme cases serve to finance **attacks** and direct action against perceived enemies. Detecting and addressing these activities with explicit economic purpose is important to curb the financial profits of actors fuelling violent extremism and to minimise the risk of (stochastic) violence and terrorism.

Post organisational violent extremism and use of borderline content

The Institute for Strategic Dialogue (ISD) shared with the EUIF a ‘**Taxonomy**’, which refers to the use of borderline content with a conceptualisation of post-organisational violent extremism.

Inspirational material		Ideological material		Instructional Material	
Content which can reinforce a violent extremist or terrorist mindset		Material which is specifically trying to further a violent extremist or terrorist worldview		Content which contains instructions on operational aspects of terrorist activity	
Non-Group	Group	Non-Group	Group	Non-Group	Group
Violent material not associated with a specific violent extremist group, designed to inspire extremism or terrorism, including copycat attacks.	Violent extremist group material designed to inspire violent extremism or terrorism.	Material not associated with a specific violent extremist group, which makes the ideological case for extremist violence.	Violent extremist group material which makes the ideological case for violence.	General instructional material providing operational guidance on carrying out acts of violence and terrorism.	Violent extremist group material providing operational guidance on conducting acts of violence and/or terrorism.

Borderline content: conceptualising post-organisational violent extremism⁸

As shown in the table, across the case studies reported by ISD in the report, they identified three broad categories for the classification of content:

- **Instructional material**, which contains guidance on operational aspects of terrorist and violent extremist activity. This includes guidance on the planning and perpetration of attacks, as well as guides on combat drills, fitness and non-violent activism such as sticker campaigning.
- **Ideological material**, which is designed to specifically further a violent extremist or terrorist world view. This includes key texts and lectures which provide the theoretical underpinning for a terrorist or violent extremist cause, and which provide explanation around why the world is a certain way.
- **Inspirational material**, designed to reinforce a violent extremist or terrorist mind-set. This includes a wide range of content which is designed to elicit a reaction or response in the radicalised mind and material intended to **provoke hatred** towards a particular group of people or **pride** and support for a particular cause. Notably, this category of content is the least well-defined in the existing literature.

7 J. Davey, M. Comerford, J. Guhl, W. Baldet, C. Colliver, *A Taxonomy for the Classification of Post-Organisational Violent Extremist & Terrorist Content*, <https://www.isdglobal.org/wp-content/uploads/2022/01/A-taxonomy-for-the-classification-of-post-organisational-terrorist-content.pdf>

8 Ibid:n

Terrorgram

ISD provided examples of content defined as part of the so-called ‘**Terrorgram**’, which refers to a network of Telegram public channels created by **white supremacists** and which became a hub for violent extremist activity. In its analysis of the content spread within Terrorgram, ISD observed not only violent content produced by groups such as Atomwaffen and The Base, but also, in a largest amount, content fitting under the category of ‘non-group affiliated’ and ‘non-violent inspirational material’, such as

- White supremacist **music and a vast-array of user-created memes** which convey racist, antisemitic and misogynist ideas or celebrate extreme right ideologues.
- A vast amount of **cultural material, and material relating to sex, gender and the family**, including historical photographs, photographs of ‘traditional’ looking beautiful women, and pictures of classical art, at times superimposed with inspirational slogans designed to reinforce a white supremacist world view, such as ‘**embrace tradition, reject modernity**’ (common features of fascist rhetoric as identified by Umberto Eco). Importantly, such apparently innocuous images were shared in the context of communities which actively advocate for extreme violence, illustrating the role which non-violent content can play in reinforcing a violent extremist mind-set, and suggesting that such material could be an indicator in certain circumstances of more concerning activity within a community.

Inspirational material			Ideological material		Instructional Material	
Content which can reinforce a violent extremist or terrorist mindset			Material which is specifically trying to further a violent extremist or terrorist worldview		Content which contains instructions on operational aspects of terrorist activity	
Violent	Non-Group	Group	Non-Group	Group	Non-Group	Group
	Oslo manifesto Christchurch manifesto Charleston manifesto Saint memes	AWD videos Azov videos The Base Serbian military videos and songs Calls for violence Glorification of historical fascists	Turner Diaries SIEGE	Hitler Codreanu GSA books	Advice for firearm construction	National Action manuals
Non-Violent	Non-Group	Group	Non-Group	Group	Non-Group	Group
	Antisemitic memes White Genocide memes Aesthetic images of European identity Fashwave White supremacist rock music Images of neo-fascist Gothic/pagan symbols		Books by Julius Evola and David Irving	Jewish Supremacyism by David Duke	Images of target buildings	

Examples of content indentified in ‘Terrorgram’ ⁹

9 Ibidem

GIFCT: defining borderline content in relation to violent extremism and processes of radicalisation¹⁰

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. Bringing borderline content to the fore of multistakeholder 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 companies¹¹ 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. Borderline content can be, and has been, conceived of in two ways.

- Academics 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 i.e., 'borderline illegal', or 'lawful but awful'.¹²
- Tech companies tend to speak about borderline content as content that **brushes up against a platform's policies for violating content i.e., 'borderline violative' but is not clearly violating a policy**.¹³ 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.¹⁴

These two categorisations for borderline content are related. It is broadly agreed that although borderline content is not always 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 recognized 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 often largely address any 'borderline illegal' TVE content through their policies.

10 GIFCT contribution to the EU Internet Forum Handbook on Borderline Content. Any feedback, questions, or comments can be emailed to Dr Erin Saltman erin@gifct.org and Micaile Hunt micaile@gifct.org.

11 GIFCT has a range of tech company members that have to reach a membership criteria in order to fully integrate with GIFCT efforts. For more on GIFCT membership and current members see: <https://gifct.org/membership/>

12 Heldt, A., *Borderline 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>

13 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/> (YouTube, 2019); 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/>

14 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>; 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>; 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/ispi/documents/reduction_ispiessayseries_jul2022.pdf.

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.¹⁵ These policy areas are both often considered '**borderline illegal**' in different circumstances and represent areas where tech companies' policies have had to become more robust in comparison with offline or real world legal guidance. 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.

These sub-themes have come up in ongoing conversations at the EU Internet Forum, within GIFCT thematic Working Groups,¹⁶ and within the Christchurch Call to action. 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. As 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.¹⁷ For any measures affecting content moderation policies and legislation should consider human rights objectives in ensuring that actions taken are legal, proportionate, and defensible.

Building off 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.¹⁸

15 See GIFCT Official Website on Membership: <https://gifct.org/membership/>

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

17 Kenney, M., *Beyond the Internet: Metis, Techne, and the Limitations of Online Artifacts for Islamist Terrorists*, Terrorism and Political Violence 22, no.2, 2010, pp. 177–97; Reynolds, S.C. and M.M. Hafez, *Social Network Analysis of German Foreign Fighters in Syria and Iraq*, Terrorism and Political Violence 31, no.4, 2017, pp. 661–86; 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; 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

18 McGuffie, K., *Applying 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/>

Borderline Content Sub-Theme Policies and Use in TVE and Non-TVE Cases¹⁹

Borderline Content Type (Policy Category)	TVE Tactics + Content Types	Non-TVE Uses or Content
- Violent Content, Graphic Content, Gore - Weaponry + Instructional Material	'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
Symbols + Slogans and Visual Indicators Associated with VE Groups	Using whistle symbols, emojis, and coded language to evade moderation efforts to remain on a platform and signal like-minded users	Academic 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 socio-political climates and scenarios that might violate a policy but are unrelated to TVE incidents or activities
- Some forms of hate leading violent extremism - Bullying, Harassment, and Threats	'Promotion of hate-based beliefs, ideologies, and discrimination. Facilitating scapegoating and false attribution of societal ills. Empowering sympathisers with misguided sense of superiority	Bullying, harassment and threats happen in a range of socio-political climates and scenarios that might violate a policy and, in some cases, be illegal, but are unrelated to TVE incidents or activities
Anti-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 socio-political 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
- Anti-Government/ Anti-EU - Anti-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) ²⁰

19 This table quotes and builds off of the work originally presented in a GNET insight by McGuffies (2021). This is a GIFCT contribution to the EU Internet Forum Handbook on Borderline Content. GIFCT contribution to the EU Internet Forum Handbook on Borderline Content. Any feedback, questions, or comments can be emailed to Dr Erin Saltman erin@gifct.org and Micalie Hunt micalie@gifct.org.
20 Table provided by GIFCT, building off of the work originally presented in GNET insight by Kris McGuffie (2021) [LINK](#)

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. Accordingly, these actors often knowingly produce content that comes close to, but does not violate, existing platform policies i.e., '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.²¹
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.²²

Experts acknowledge it is necessary to provide parameters to the term.²³ 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.²⁴

EIF review of Analysis and definitions of Dehumanisation

In her study called 'Heroes and Scapegoats', researcher and data analyst Lisa Kaati identifies targeted toxic language, dehumanisation and conspiracy theories as means to create conceptions about groups of people as enemies: threatening, malicious and of lesser value. These conceptions may serve to justify contempt, discrimination and violence.²⁵

As observed by Kaati, the term 'hate speech' is often used to cover different forms of expressions that spread, incite, promote or justify hatred, violence, and discrimination against a person or group of persons (and, in this case it is illegal in EU Member States). However, while most definitions of hate speech share some common elements, alternative terminology such as 'abusive language', 'toxic language', and 'dangerous speech' have been introduced to either broaden or to narrow the definition. The term 'dangerous speech' is used to describe communication that inspires violence and rhetorical techniques such as dehumanisation, protection of an in-group, and threats towards the purity of a group.²⁶

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

22. Allchorn, W., *Turning Back to Biologised Racism: A Content Analysis of Patriotic Alternative 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/>

23. Rogers, E., *The Role of User Agency in the Algorithmic Amplification 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/>

24. J. Davey, M. Cornerford, J. Guhl, W. Baldet, C. Colliver, *A Taxonomy for the Classification of Post-Organisational Violent Extremist & Terrorist Content*, <https://www.iscglobal.org/wp-content/uploads/2022/01/A-taxonomy-for-the-classification-of-post-organisational-terrorist-content.pdf>

25. L. Kaati, *Heroes and Scapegoats. Right-wing extremism in digital environment*, 2021 European Commission's Directorate-General Justice and Consumers.

26. L. Kaati, *Heroes and Scapegoats. Right-wing extremism in digital environment*, 2021 European Commission's Directorate-General Justice and Consumers.

While providing examples and trends on the use of toxic language by violent extremists, Kaati explains that the reason for using different terminologies is the need to differentiate illegal hate speech from legal expressions of hate or aggression toward certain groups, while still acknowledging that the latter may also be harmful.

In the context of prevention of terrorism and violent extremism, a key partner of the European Commission is the Christchurch Call to Action's network. Especially on borderline content leading to violent extremism, the EU Internet Forum has benefited from the cooperation with its network and have exchanged on a regular basis on best practices, definitions and policy work carried out on this subject matter.²⁷

In 2022, the **Christchurch Call Advisory Network (CCAN)**²⁸ led work to understand Members' responses to **dehumanising speech** and discourse. As observed by CCAN and Christchurch partners in recent meetings, terrorist attacks such as the one in Christchurch and Buffalo in 2022, as well as the ISIS genocides of Yazidi people, were all preceded by the dehumanisation of an outgroup. Dehumanising speech or language includes describing a group based on a protected attribute as biologically subhuman ('cockroaches,' 'microbes,' 'parasites,' 'yellow ants'), mechanically inhuman ('logs,' 'packages,' 'enemy morale'), or supernaturally alien ('devils,' 'Satan,' 'demons'). Dehumanising discourse portrays a protected group as polluting, despoiling or debilitating society or presenting a powerful menace or existential threat to society (e.g., conspiracy theories like the Great Replacement Theory).

In March 2022, CCAN distributed a request for information to all members of the Christchurch Call, including tech companies and governments. Based on the replies, CCAN suggested measures to reduce the spread of dehumanising content online, as part of the Call's commitments.

According to CCAN, dehumanisation provides a valuable framework for policy for the following reasons

- Dehumanisation is distinct from hate speech because it aims to lower an audience's moral reflexes towards a particular group by removing them from the human family.
- It is more readily apparent and definable as conduct than concepts of disinformation, hate speech or extremism on their own.
- It provides a lasting framework that can respond to changes in discourse and targeted groups over time. It explains how a range of groups based on race, religion (including no religion), gender, gender identity, disability and other protected attributes are dehumanised through hate speech and cumulative discourse.

From the responses received, **members of the Call (except for Twitter) did not have explicit policies** or responses to dehumanising speech. Under half of the respondents had civil laws that could penalise dehumanising language or discourse provided it met specific thresholds in their legislation. Of those examples, the burden remained largely with the community to bring forward complaints.

27 On 15 May 2019, two months to the day after the terrorist attack on two mosques in Christchurch, New Zealand, which killed 51 people and injured 50, New Zealand Prime Minister Jacinda Ardern and French President Emmanuel Macron brought together Heads of State and Government and leaders from the technology sector to adopt the Christchurch Call. The European Commission has been an active partner of the call since its creation. The Christchurch Call is now a community of over 120 governments, online service providers, and civil society organisations acting together to eliminate terrorist and violent extremist content online.

28 Full report on the CCAN website: <https://christchurchcall.network/>

All respondents had **criminal laws** that could **penalise dehumanising language or discourse if it met the threshold** in their legislation. However, CCAN observed great variations in the thresholds. Some laws only apply to direct attacks on individuals online rather than direct or cumulative attacks on group identities. Dehumanisation through discourse, such as disinformation, may not meet the harm threshold. CCAN therefore asked that the Call create space to discuss policy responses and definitions for dehumanising speech and discourse. This would elevate the perspective of terrorism victims and strengthen the prevention-focused efforts of the Call.

The Call may wish to recognise:

- The **serial or systematic dehumanisation of an out-group** identified on the basis of a protected characteristic is a form of violence, an attribute of TVEC and a driver of violence.
- Portraying groups through curated information as **polluting, despoiling or debilitating society** or as an existential threat to society is a powerful form of dehumanisation.
- Dehumanisation creates **risk for targeted groups, society, and democracies**.
Levers for enforcement need to be carefully considered in any response.

Recommendations from CCAN:

- The **burden of enforcement** should not remain on targeted communities.
- The **response** should **protect** communities (not only individuals) and cover cumulative harm.
- The response should prioritise **non-carceral approaches** that are fit for purpose.
- Relevant decision-makers must have enough **independence** and scope to consider all contributors to dehumanisation.
- **Human rights diligence** by civil society needs to be encouraged.

Categories of borderline content in relation to radicalisation and violent extremism

Racist, anti-migrants and antisemitic content

29



Member States reported about the increase in antisemitic messages and posts identified online and called for better moderation of this type of content, even when the content is illegal hate speech, but its illegality is sometimes difficult to assess or to relate to violent extremism and terrorism. Messages against both Jewish people and the state of Israel were detected, often with dehumanising, hateful messages supported by conspiracy theories against Jewish targets. An example of disinformation narrative targeting Jewish people is the denial of the Holocaust or its so-called ‘exaggeration’ by Jewish run media. Jewish people are often used as ‘scapegoats’ and blamed for all major crisis, including the COVID-19 pandemic and the war against Ukraine. A number of Member States pointed at V Kontakte as platform where many anti-Semitic posts were identified.

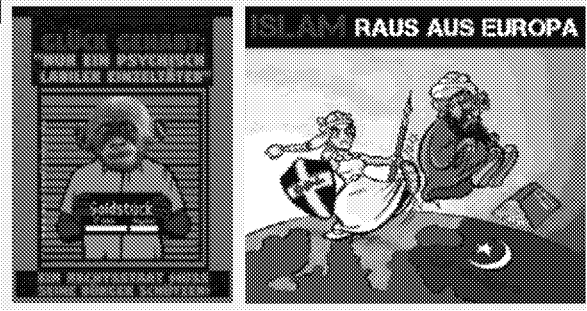
30



The use of disinformation and manipulation techniques is also popular among anti-migrant users. Recently, disinformation against migrants was used to legitimise violence and killing of South American migrants. A woman called Romana Didulo, who self-declared herself as Queen of Canada, put out a message to her over 48,000 Telegram followers telling them that ‘illegal migrants’ trying to cross the United States to then enter into the country ‘should be shot on sight’.³¹

Humour is often used by violent extremists to dehumanise or express hate against migrants, especially those of Muslims background, and political opponents. It makes the detection and moderation of such content by tech platforms more challenging.

32



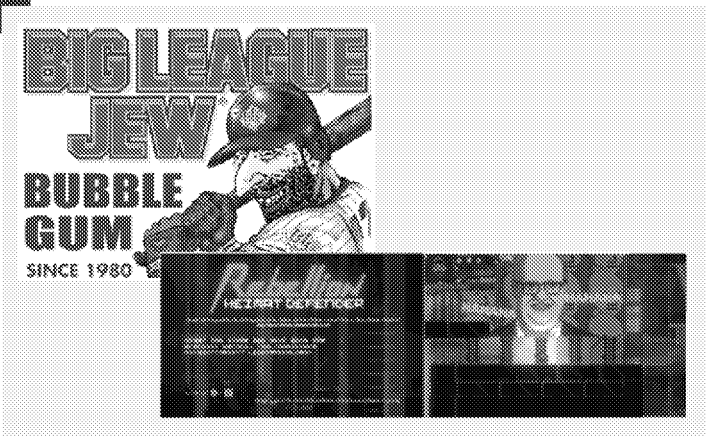
An EU Member State provided an analysis of systemically screened and categorized extreme right-wing claims, collected between 2019 and 2020, and proves the movement’s ability to monitor and frame political and discursive opportunities and place ideas in online communities. The Member State informed about interactive formats such as music, lifestyle, and video games that have the potential to influence

29 Image provided by Romanian authorities.
30 Image from Vice magazine’s article [LINK](#) Pic quote: screenshot from QAnon Queen Romana Didulo’s Telegram.
31 Article on [antihate.ca](#) [LINK](#)
32 Image provided by German authorities for the scope of drafting this handbook.

younger audiences in a negative way. These formats ensure visibility with attention-grabbing and entertainment-focused content, and become digital repertoires of contention for individuals or groups that oppose a public decision they consider unjust or threatening. Results show that it is common for topics selected by extremist groups to be in the grey zone of illegal and legal content.

The German project ‘Control Propaganda Online: Development of Crime Prevention Tools to Curb Extremist Propaganda and Hate Messages on the Internet’ funded by the EU-ISF provided the following examples of borderline content spread by violent extremists to radicalise users.

33 34



Here are some examples of posts against migrants, Muslims and Jewish people identified as borderline content in relation to violent extremism and radicalisation by German authorities.

35



‘Remove Kebab’ is a slogan used in Islamophobic memes online, which calls for deportation or even killing of Muslim people. Its origin comes from a Serbian music video containing anti-Muslim propaganda during the Yugoslavian civil war. The slogan and the image of soldiers playing the accordion are used in memes and in quotes on online profiles. Christchurch attacker, Brenton Tarrant, wrote ‘Kebab Remover’ on the barrel of his rifle as a message to the extreme right-wing image board scene.

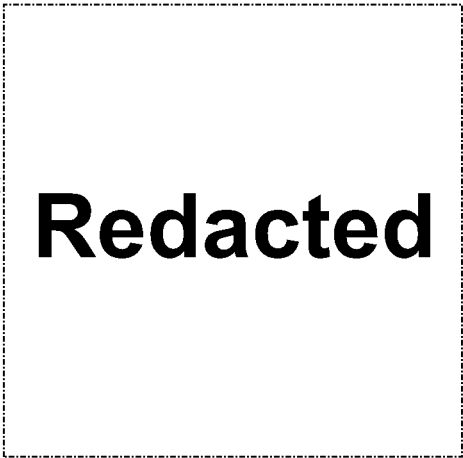
36



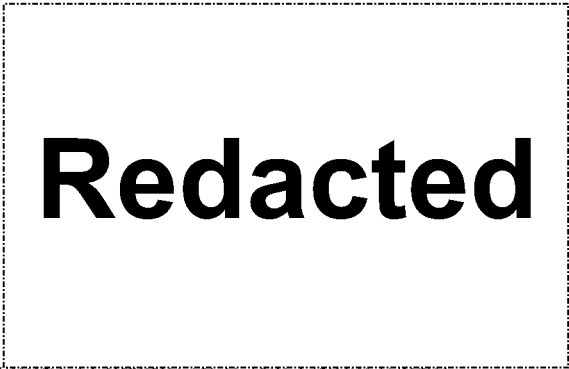
German authorities also identified the use of **Echo/Echoing** and **triple brackets** as an **anti-Semitic** meme, which consists in a word being between the three brackets branding what is between them as Jewish. For example, the sentence ‘You know (((whom))) I mean...’ implies that the antisemitic author is speaking of Jews. In the hereby examples, purely textual and a graphic variants of the meme are used. In the first one, the triple parenthesis is applied within an antisemitic post on Twitter to denigrate Angela Merkel as a ‘Jewish conspirator’. The example image below also uses the anti-Semitic Happy Merchant meme.

33 Image provided by German authorities for the scope of drafting this handbook.
34 Ibidem
35 Ibidem
36 Ibidem

Countryballs/Polandball are very widespread memes mainly used as form of political caricature. As shown in the accompanying images, the drawings consist on balls representing different countries with stylised faces. This kind of meme is often used to comment on current political events and does not necessarily have violent right-wing extremist references. In this specific example, the right wing extremist group Junge Alternative NRW depicts the negotiations on the refugee agreement between the Maghreb countries and Germany as a tug-of-war between 'Germanyball' and 'Tunisiaball' over a detonated bomb with 'Islamismus' written on it. In this case, Tunisian migrants are deceptfully defined as terrorists.

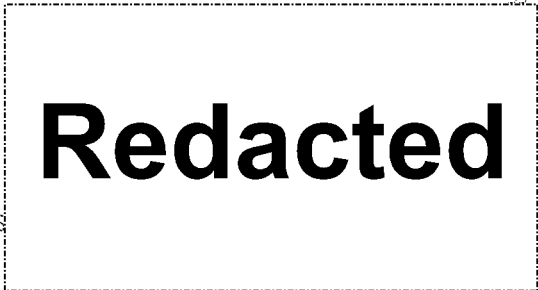


Borderline content targeting Jewish people, Muslims and migrants can also be found represented in merchandising, often sold to raise money for terrorists and violent extremists. **'ZOG'** is the acronym for 'Zionist Occupied Government'. It refer to a far-right conspiracy theory reflecting the idea that the government is controlled by Jews.



'JWO' is the acronym for 'Jewish World Order', the antisemitic version of the New World Order conspiracy theory, claiming that the single world government will be lead by Jewish people.

'NOW' is the acronym for 'New World Order' a conspiracy theory that argues that a shadow elite is trying to implement a totalitarian world government. During the Covid-19 pandemic, the conspiracy theory gained new supporters claiming that this 'world government' is achieved through a manufactured crisis with the intention to exert control over citizens.



Merchandise that includes the following terms or slogans (not exhaustive) should be analysed closely to assess a link to violent extremist ideology or conspiracy theories:

- No white guilt
- Save Bees not Refugees

37 Ibidem

38 Source for the images: T. Squirrell, C. Martiny, *Profiting from hate*, ISD, 2022 p. 14-17

39 Source for the images: T. Squirrell, C. Martiny, *Profiting from hate*, ISD, 2022 p. 14-17

Covid-19 crisis-related borderline content

40

Redacted

In a number of EU Member States, disinformation on vaccines and the role played by governments in preventing the spread of COVID-19 were disseminated by extremist groups to reinforce their propaganda against public figures, governments, mainstream media, as well as Jewish people and migrants, who were often blamed for the crisis.

Conspiracy theories claimed the vaccines were going to kill an entire generation, or that people were being enslaved and forced to vaccinate against their will.

In some cases, violence perpetrated in the West against authorities was incited and glorified by violent extremist online.

The **Radicalisation Awareness Network (RAN)**⁴¹ reported on how the COVID-19 pandemic and measures against its spread and their impact on populations, created an opportunity for new dynamics of development of violent right-wing extremism (VRWE). Utilisation of the 'corona-crisis' became typical of various forms of violent extremism and terrorism in a global scope (including violent religious extremism). The specific threat of VRWE is characterised by various forms of **violent behaviour** (so-called corona hate crimes, violent demonstrations and riots, discussions about terrorist attacks, etc.) and by the interconnection of **VRWE with a broader spectrum of mass protests** over responses to the COVID-19 pandemic and with mass propagandist campaigns on new social media channels.

The most important challenge connected with VRWE in the EU during the corona-crisis is connected with mass violent demonstrations and unrest. With the aim to mobilise people on the streets, right-wing extremists/violent right-wing extremists use the **mass spreading of fake news and conspiracy narratives** in combination with capitalising on the frustration and depression of many people (caused by the impact of anti-COVID measures on their daily lives). The core of the conspiracy narratives used remains relatively consistent. It is adaptable to new trends and events in the pandemic and the measures taken to prevent these narratives from spreading. These narratives are not typical only of RWE/VRWE, but are also spread within left-wing extremist and religious extremist milieus. Still, the role of violent right-wing extremists in the EU is the strongest in comparison with other variants of extremism.

The conspiracy narratives widespread in the EU are connected with the spread of conspiracies on a more global scale (including QAnon, with origin in the United States). They are supported by anti-EU actors, mostly by Russian governmental and pro-Kremlin forces, with the aim to undermine the authority of the EU and its Member States. The most important RWE conspiracies related to COVID-19 (19) are:

- COVID-19 is a **Chinese weapon** against the West.
- COVID-19 was created as a tool of global elites and the **'deep state'** (represented by Bill Gates, George Soros, 'Brussels elite', etc.), or even the **'Jewish elite'** widespread with the aim to
 - **Earn money from masks**⁴² and vaccination (also the big pharmaceutical industry should be involved in this conspiracy),
 - **Eliminate the Western population** (and to replace 'traditional citizens' with migrants), or take control over free citizens (microchips in vaccination).
- The use of the **5G transmission** towers narrative for the spread of COVID-19 (20) can be added as a subsidiary conspiracy. Anti-pandemic measures are misused by governments and by globalists to **destroy small entrepreneurs** and to create a dependent population.

40 Images provided by Italian authorities for drafting this handbook.

41 European Commission's RAN paper, *Capitalising on crisis*, 2021

https://home-affairs.ec.europa.eu/system/files/2022-02/ran_capitalising_crisis_how_vrwe_exploit_covid-19_pandemic_082021_en.pdf

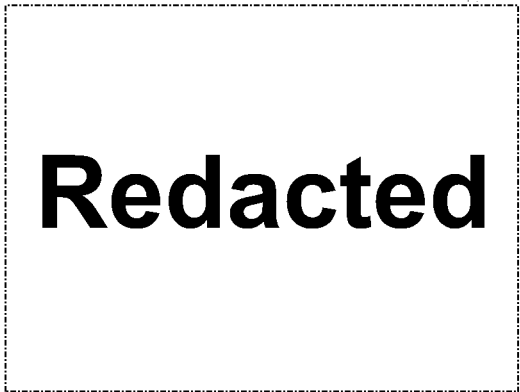
42 Image provided by Romanian authorities.

How conspiracy theories, disinformation and dehumanisation can lead to violence

The most important violent right-wing extremist activities in the EU connected with the COVID-19 pandemic were in 2020 and 2021 and consisted in mass violent demonstrations and unrest organised by violent extremists with the **intention of gaining media attention**, mobilising their own supporters and strengthening their political position. Speeches, mottos and arguments supporting violence included elements of **fake news and conspiracy narratives**.⁴³

Violence plays a symbolic role in the following ways:

- For violent extremists to show how through violent activities they are able to **recruit and radicalise common people** who did not previously belong to extremist milieus.
- The capability to organise violent activities **strengthens the self-confidence** of far-right militants.
- For some young people, **videos with clashes** are an inspiration for similar behaviour.
- Serious **threats towards politicians and experts** who were considered as responsible for governmental measures (or who publicly defended these measures) were posted online, including by people who did not belong to extremist milieus, but were successfully influenced by narratives spread by violent extremists.

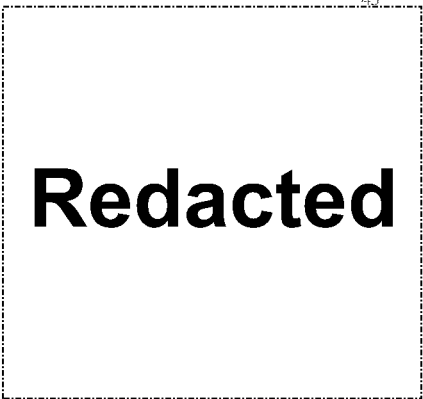


From the perspective of VRWE development, the era of the corona-crisis in 2020/2021 can be characterised by:

- Specific **hate crimes and violent mass protests** related to COVID-19 and countermeasures.
- Temporary dominance of other topics in some countries for local VRWE scenes (BLM protests, clashes of violent right-wing extremists in Poland with opponents of the anti-abortion law, anti-police protests in France, etc.).
- Deepening of global interconnection of chats and webs with **conspiracy narratives, fake news and hate speech** with the potential to radicalise violent right-wing extremists and terrorists in the future.

Member States provided examples of borderline content on COVID-19 used by violent extremists to radicalise and recruit new cadres.

Various excerpts from popular series and films are regarded as particularly iconic 'cult moments' within the network culture and are also used as such in memes. Even within the **online subculture of the Alt-Right**, these iconic excerpts are often reused and manipulated to send extremist political messages and propaganda. The sample image is from the US sitcom 'Friends' and shows one of the main characters wearing a **skull half mask**, which refers to actors of the violent right-wing extremist white power movement **Atomwaffendivision**. The caption of the meme refers to extremist political violence. The skull mask was also seen at recent violent protests and is a nod to extremist underground meme communities.



43 European Commission's RAN report, *Capitalising on Crises, How VRWEs Exploit the COVID-19 Pandemic and Lessons for P/CVE* https://home-affairs.ec.europa.eu/system/files/2022-02/ran_capitalising_crises_how_vrwe_exploit_covid-19_pandemic_082021_en.pdf
44 Image provided by Romanian authorities.
45 Image provided by German authorities for the scope of establishing this handbook.

Other EU Member States observed the involvement of **new actors**, mainly with no previous links to the extremist spectrum, in anti-vaccination and anti-COVID measures disinformation campaigns which led to violence. These actors can be linked to the new wave of anti-constitutional extremism observed in several EU Member States active on several social media platform and communication services.

How right wing extremists use disinformation campaigns

In the following table, the Radicalisation Awareness Network provides information on the major conspiracy theories developed and disseminated by right wing extremists to sow mistrust and, in some cases, incite to violence against democratic governments and institutions, mainstream media, political opponents, as well as against the members of specific cultural and religious milieus and migrants, which are deceptively described as 'elites' or 'enemies'.

46

Radicalisation Awareness Network RAN		CONCLUSION PAPER IMPACT OF CONSPIRACY NARRATIVES ON VIOLENT RWE & LWE NARRATIVES Page 4 of 6	
		THE GREAT REPLACEMENT	QANON
1	G [GOAL]	• Civil war; • Polarisation; • Divide and conquer; • Dehumanise; • Sense of identity; • Keep the "traditional" community; • White ethnostate.	• Defeat evil; • To provide psychological comfort to their confusion in a complex world / understanding confusing reality; • Restore an enchanted world; • Trump presidency.
2	A [AUDIENCE]	• People who experience a sense of loss / grieved individuals; • Working class; • The supposed "elites"; • Conservatives.	• Alt-right / patriots; • Ordinary people (mothers and fathers; men and women); • People who are wrestling with the complexities of the world: searching for meaning.
3	M [MESSAGE]	• "We are under attack"; • "It is now or never" / "We have to defend ourselves"; • "Destruction of traditional family is reason for decline of EU populations"; • "Immigrants are invaders"; • "Mistrust of mainstream media".	• "The elite ruling is the evil" / "Everything you know is a lie"; • "Good, ordinary people are victims"; • "Your situation is the product of intentional plans by others"; • "Your liberty is under threat".
4	M [MESSENGER]	• Influencers / self-proclaimed political figures / academics / journalists; • Terrorists / extremists; • Religious leaders; • Alternative media; • Social media bots; • Accelerationists.	• Q; • Alternative media; • Individuals claiming to be "red-pilled"; • Influencers; • Echo chambers / every follower is a messenger.
5	M [MEDIUM]	• Fringe platforms: Gab, 4chan, 8kun; • Mainstream platforms: TikTok, Reddit, WhatsApp, Telegram; • Print media: stickers and posters; • Live streams / video games; • Real-life events.	• Large social media platforms; • Fringe social media platforms; • Memes / live streams; • The Dark Web.
6	A [ACTION]	• Political action: Change the system from the inside / overthrow of government; • Raise awareness / protesting; • Promote conflict of truth / violence; • Red-pilling.	• Spread the message; • Protest / riot; • Prepare for violence against immoral other / prepare for martial law.

46 European Commission's RAN report, *Capitalising on Crises. How V/RWEs Exploit the COVID-19 Pandemic and Lessons for P/CVE*
https://home-affairs.ec.europa.eu/system/files/2022-02/ran_capitalising_crises_how_vrwe_exploit_covid-19_pandemic_082021_en.pdf

The focus of the table is on QAnon and the Great Replacement. It provides insights into the objectives/goals, the types of message sent, the target audience, the platforms and media used to spread the messages, and the type of actions that violent extremists spreading these conspiracy theories are calling for.

Amplification tactics used by violent extremists: the superspreader of borderline content leading to radicalisation and violent extremism

Online actors responsible for the widespread distribution of harmful content are often defined as **superspreaders**. Often widespread propaganda campaigns disseminating malicious content are triggered by a **small group of users**. An example of superspreader activities is discussed in an article published by Canadian McGill University⁴⁷ in Montreal, which refers to a report by the Center for Countering Digital Hate (CCDH), 'The Disinformation Dozen'. The report reveals how **dozen misguided influencers spread most of the Anti-Vaccination Content on Social Media**. Its main takeaway is that two-thirds of anti-vaccine content shared or posted on Facebook and Twitter between February 1 and March 16, 2021, can be attributed to just twelve individuals. Twelve. Let that sink in.

Examples of the types of deceptive messages sent out by these superspreaders include claiming that vaccines make women infertile or that they have killed more people than the disease itself.

This modern anti-vaccination movement is led by **well-financed influencers** who have accumulated a sizeable following on social media platforms, where, in some cases, fear spreads more easily than facts and nuance. Among them are anti-vaccine activists, as well as alternative health entrepreneurs and physicians. Some run multiple accounts across the different platforms and often promote 'natural health' and even sell natural supplements, beauty products, pet supplies or books.

Tactics used by anti-vax superspreaders

Anti-vax superspreaders try to innovate their message and sometimes skirt platforms' rules by using secret codes. For instance, **instead of saying 'vaccine,' they may, in a video, hold up the V sign** with their fingers and say, 'If you're around someone who has been' — hold up V sign — 'you know, X might happen to you'. In other cases, they falsely link a famous person's death to the fact that the celebrity was vaccinated days or weeks earlier. Some of the influencers even use a strategy called **'tried and true'**. With this tactic, anti-vaccine influencers try to promote some sort of **research and data to foster any narrative** they have chosen. They do not seek a logic or scientific evidence, but evidence that can justify their assumptions.

Borderline groups

The Italian anti-vax and anti-green group 'borderline movement'⁴⁸ called **'Guerrieri V_V' (alive warriors), active online with the label 'V_V'**, was intercepted and investigated by the Italian police on online platforms such as Telegram. The group was accused of disseminating **violent propaganda against political figures** on social media platforms in a coordinated manner. Violent attacks and language were used by the group against people expressing opinions in favour of the vaccination campaigns and 5G. Threats and aggressions were glorified and amplified on platforms, especially on Telegram, but were not automatically removed by all platforms, because they did not meet the thresholds to be considered harmful or illegal.

47. Jarry M.sc., *A Dozen Misguided Influencers Spread Most of the Anti-Vaccination Content on Social Media*, McGill University, Montreal: <https://www.mcgill.ca/oss/article/covid-19-health/dozen-misguided-influencers-spread-most-anti-vaccination-content-social-media>

48. As described by the Italian National Police, which reported of violent extremist groups that they have labelled as borderline because they are not proscribed nor banned.

The 'V_V' movement, was investigated by the Italian National Police for **intruding in an online meeting** held by the local public administration of Trieste on 'GoToMeeting' with images of t-shirts bearing the logo of 'V_V' and other defamatory and precatory sentences with an anti-vax and anti-Covid measures orientation. A group called '**Blocco Est Europa**' (**East Europe block**) was also intercepted online by the Italian police on Telegram for sharing racist and violent content with messages referring to racial hate. Antisemitic messages and glorification of Hitler were intercepted, alongside hateful posts against public institutions and law enforcement and gender-based hate speech. Videos of suicidal and murdered women, especially young ones were identified. Training on the use of weapons were detected on the groups' online channels and profiles. Not all this content was treated as illegal or in relation to violent extremism and terrorism.

Lastly, the NGO ISD, in its report 'Profiting from hate', provided examples of how **borderline content** referring to the COVID-19 pandemic and vaccines was spread online **to radicalise and raise money** for violent extremist groups. Here are some visual examples bearing COVID-19 disinformation.

49



Pro-Kremlin borderline content

50



Since the breakout of the war of aggression against Ukraine in 2022, violent extremist groups have been using the conflict to radicalise and recruit users online. As reported by Member States, borderline content, mainly disinformation, is being used by pro-Kremlin users for disrupting society and fostering anti-Ukraine sentiments.

Some EU Member States have informed about the activities of so-called 'borderline' groups, often defined as **'anti-system'** violent protesters and conspiracists with links to violent extremism, who were spreading anti-vax narratives to incite violence against governments. Law enforcement agencies and experts revealed how these same **groups are now disseminating pro-Kremlin disinformation propaganda** often with reference to **xenophobic-racial** content, to sow hate and violence against specific targets (the EU, the West, NATO, Jewish people, Ukrainian migrants).

49 Source for the images: T. Squirell, C. Martiny, *Profiting from hate*, ISD, 2022 p. 13
50 Image provided by Czech Republic's authorities for the establishment of this handbook.

Examples of deceptive content identified by Member States usually refer to the following allegations:

- **NATO** (and EU) spread(s) propaganda
- Russian invasion is **not an invasion**
- Russia's so-called „special operation' is an act of **peace**
- Russia does not attack cities and civilians
- Ukrainians are **fascists**
- Ukraine's president is controlled by evil **Zionists**

Fakes and disinformation by 'borderline groups'

51



The use of **fake photos and videos**, in which the either the existence of a war is denied, or where Ukraine is described as the initiator of the conflict and blamed for war crimes. In these accompanying images, pro-Russia users claim scenes portraying Ukrainian victims are fake.

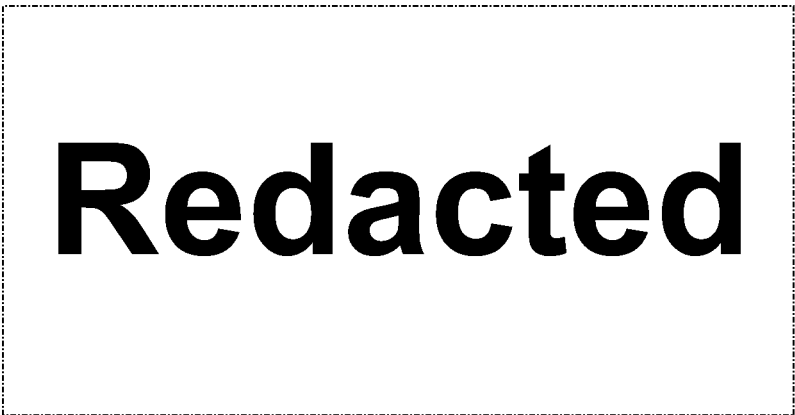
Conspiracy theories and fakes were identified on a number of platforms (such as Vkontakte, Telegram, TikTok, Facebook) and they usually refer to:

- **Revisionist ideas** disseminated to sow anti-West sentiments
- Disinformation and conspiracy theories about **Ukrainian refugees**
- Main **anti-European narratives: EU/NATO** will drag us into the conflict

Eastern-European EU Member States reported of the activities of pro-Kremlin influencers spreading disinformation about the war of aggression against Ukraine.

Influencers and dedicated websites identified in the Czech Republic

52 53



An example of a **pro-Kremlin influencer** active in the Czech Republic is Nela Lisková, self-proclaimed ambassador of separatist Donetsk People's Republic in the Czech Republic, who was identified for posting disinformation about:

- Alleged Ukrainian fascists.
- EU and USA acting as nefarious entities against Russia.
- The UN ignoring the so-called genocide of Russian minority in Ukraine by the Ukrainian Government.

51. Image provided by Italian authorities for the establishment of this handbook.
52. Image from Facebook's post provided by Czech Republic's authorities for the establishment of this handbook
https://www.facebook.com/groups/pratele.ruska.v.cr/posts/2136266243204834/?comment_id=2136439646520827
53. Image from Facebook's post provided by Czech Republic's authorities for the establishment of this handbook
<https://www.facebook.com/tadesco.org/posts/311289684318950>

Examples of **websites** spreading pro-Kremlin disinformation narratives in the Czech Republic, which are potentially leading towards violent extremism are:

Protiproud, with content related to common disinformation narratives, such as:

- The Russian invasion is an act of peace.
- Ukraine is a boil which must be removed.

54



Tadesco, which has spread fake news about alleged **US biolabs** financed by US MoD on Ukraine.

Former politicians are spreading narratives on alleged Ukrainian fascists using banned phosphorus ammunition in Kyiv. The same politician circumvented the EU ban on Russia Today (RT) and shared pro-Kremlin unverified information coming from RT on social media platforms.

Romanian authorities confirmed that borderline content related to the Ukrainian conflict identified in Romania usually refers to revisionist rhetoric and anti-European/anti-immigration narratives.

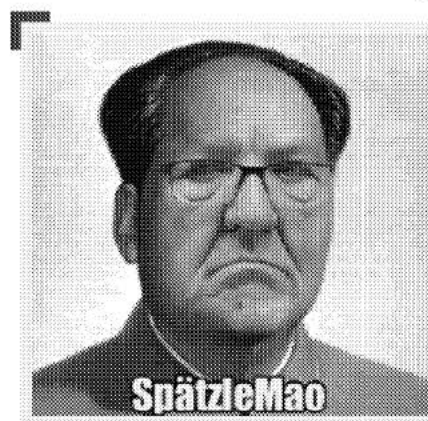
Romanian authorities identified a series of revisionist ideas spread by pro-Kremlin extremists, especially on Facebook, such as:

- 'Ukraine is an **artificial** state'.
- 'Lost territories like Ukraine must be **reannexed**'.
- 'Ukrainian **migrants** are illegal'.
- '**EU and NATO** have created the conflict and will drag Eastern EU countries into it'.

54 Image from Twitter's post provided by Czech Republic's authorities for the establishment of this handbook
https://twitter.com/ubomir_voiny/status/1497007199884828674

Under the **German Project** 'Control Propaganda Online: Development of Crime Prevention Tools to Curb Extremist Propaganda and Hate Messages on the Internet' **online posts linked to the war against Ukraine** that could fall under the definition of borderline content were identified.

An example was provided which was shared on Telegram in July 2022. The title of the post in German is: 'Der Krieg gegen Russland wird unter der Dusche entschieden' (the war against Russia will be decided 'under the shower'). A text is added to the picture of the president of Baden-Württemberg region, MP Kretschmann, reading: 'If you shower for two minutes instead of eleven, you save 80 percent of the energy you need for the shower. [...] Now it really depends on everyone'. A comment reads: 'The Spätzlemao' (referring to Spaetzle, a dish typical of Baden-Württemberg) – a comparison between Chinese dictator Mao and Winfried Kretschmann.



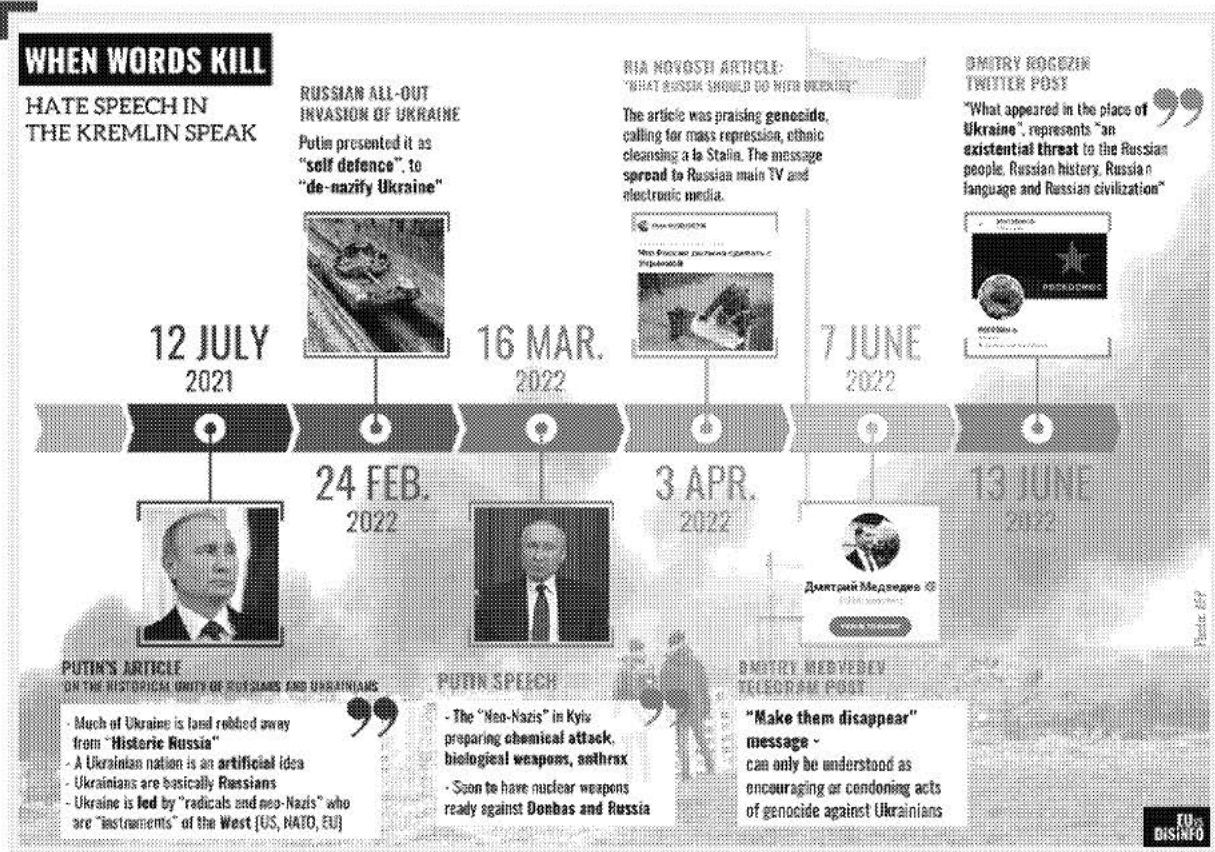
Hate speech has two 'traveling companions' – disinformation and media manipulation. Russia's war against Ukraine demonstrates the **deadly effect of hate speech**, as it has served to dehumanise the opponent, in this case the legitimate, elected government in Kyiv and the wider Ukrainian population.

Redacted

Once the foe is dehumanised, soldiers on the battlefield do not fight another person like you and me, but rather a lower-ranking group. A EUvsDisinfo analysis carried out by the EU External Action Service (EEAS) in the EU shows how a one-time Russian president Dmitry Medvedev used Telegram to spread **hate speech** in the form of claims that 'all Ukrainians should be **wiped from the face of the earth**'. And apparently, it is not the first time that Mr Medvedev has riled the followers of this Telegram channel amounting to incitement of violence and justification of war crimes.

55 Image provided by German authorities for the establishment of this handbook.

56 EUvsDisinfo, Kremlin Hate Speech Incites War Crimes in Ukraine, June 09, 2022 <https://euvsdisinfo.eu/kremlin-hate-speech-incites-war-crimes-in-ukraine/>



This outburst represents the summation of a narrative of **dehumanising and vilifying** all Ukrainians, equating them to Nazis and calling for their eradication in a manner that can only be described as genocidal.⁵⁷ Language matters and words spoken with such unbridled hatred can lead to very real and very tragic consequences for the people in Ukraine. And those who dare to speak such words must be held to account in the same way as those who pull the trigger.

EUvsDisinfo reminds users of how Russian state-controlled disinformation outlets have always used the '**Russophobia**' argument to explain away any Western criticisms or counteractions. The pro-Kremlin disinformation ecosystem in 2022 revamped the **#StopHatingRussians** social media campaign, slinging baseless accusations of Western attempts to cancel Russian culture.

⁵⁷ EUvsDisinfo, Kremlin Hate Speech Incites War Crimes in Ukraine, June 09, 2022 <https://euvsdisinfo.eu/kremlin-hate-speech-incites-war-crimes-in-ukraine/>

⁵⁸ EUvsDisinfo, Kremlin Hate Speech Incites War Crimes in Ukraine, June 09, 2022 <https://euvsdisinfo.eu/kremlin-hate-speech-incites-war-crimes-in-ukraine/>

The use of irony and memes

Although the limits of humour can be debated in many different ways and consequently it is difficult to find a consensus as to where those limits lie, the use of irony and memes in the online dimension is nowadays matter of concern. This can be explained with the fact that any meme can be exploited by entities or individuals interested in disseminating hate speech or extremist **code messages** that cannot easily be labelled as violent extremist.

Memes stir up **emotion** and generate strong reactions. Usually, creators of memes use **neologisms, metaphors, slogans and fixed phrasal expressions**. Abbreviations as part of informal language are also employed in memes to convey meaning in as little words symbols as possible.⁶⁰ They encourage **attention-grabbing** over slow-burning systematic, contextualised thinking. Nevertheless, borderline content and memes have been used by right-wing extremists to **mask** the content or the championing of some biases with ridiculous representation. The aim is to present their views or ideology in a format that eludes the policies/ rules imposed by social media platforms, communication services and the legislative framework. They have promoting ideas through 'memes' that seem inoffensive but that in fact spreads a **xenophobic, homophobic, ethnic/religious** intolerant ideas.

A study led by the Italian Police, discovered that usually these people use confidential communications platforms and restricted chats (so-called satellites, regional, etc.). Moreover, they utilise sophisticated tools for **account anonymisation, fake news** and borderline content to avoid being pursued.⁶¹

Furthermore, research in Romania revealed that many **memes are being created as 'jokes'**, but the images contain extremist beliefs and are aimed to spread hate speech. Usually encountered on Instagram and Facebook, this type of content is popular among teenagers and young adults as the following examples shows.

Historical footage of armed forces combined with racist jokes is a popular Internet meme among right-wing extremists, but obviously not all memes with pictures of the army must be considered as right-wing extremist. They must always be read in context. The use of these memes by violent extremists is a subtle attempt to **provoke humour and shift a socially acceptable discourse towards extremist right wing** contexts. The example image suggests that one soldier tells the other to load his machine gun with Jew memes - a request among right-wing trolls to flood the current forum discussion with antisemitic jokes.

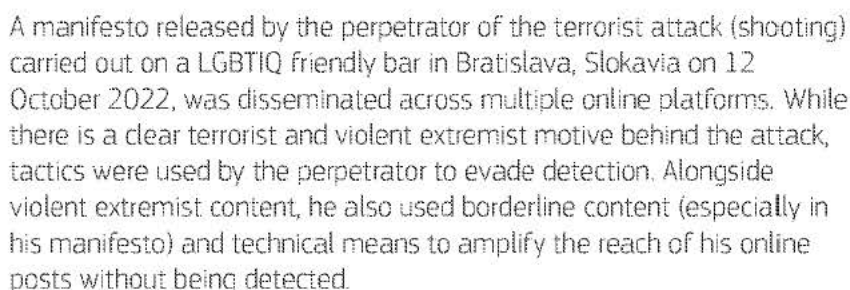


59 Image provided by Romanian authorities for the establishment of this handbook.

60 S. Fubara Joshua, A Pragmatic Analysis of the Discourse of Humour and Irony in Selected Memes on Social Media, University of Port Harcourt [LINK](#)

61 This information comes from the presentation given by Italy in the EU Internet Forum workshop on borderline content held on 29 September 2022 in Brussels.

62



The content posted by the perpetrator is defined by Tech Against Terrorism as ‘heavily antisemitic, racist and homophobic rhetoric’. In the context of **antisemitism**, Tech Against Terrorism points at content ‘that incites violence against what the user refers to as ‘the **jew machine**,’ which is a keyword probably referring to the tech industry’. The hashtags used in the perpetrator’s tweet are: #bratislava #hatecrime #gaybar

Spread of content

Redacted

64 Link to the European Commission's official page LINK

The attacker's second period of tweeting activity included hateful and extremists memes, tweets using antisemitic and racist slurs, and positive comments about other far-right terrorist attacks.

From August 2022, more unambiguously hateful and violent content was posted alongside pictures of himself in front of the LGBTIQ bar where he would later carry out the attack and the **house of the Slovak Prime Minister**. According to his manifesto and communication on 4chan, these pictures may have been some sort of **memorabilia** of him planning the attack. While the content posted during this time began as mainly antisemitic and anti-Black hate speech, it expanded in September to include more **anti-LGBTIQ content**.

65



News of the attack was quickly greeted with statements **supporting far-right violent extremists and terrorist networks**, particularly on **Telegram**, where in some instances the perpetrator was described as a '**new Saint**'. In some other online spaces popular among far-right extremists, the response was negative, with users questioning the motive and operational purpose of the attack. TAT analysis of content on the perpetrator's Twitter profile since September 2022 showed frequent references to neo-Nazism, alongside **heavily antisemitic, racist and homophobic rhetoric**.

Shortly after the attack took place, Tech Against Terrorism identified 32 unique URLs hosting the manifesto online, spread across 17 different platforms, categorised as messaging, forum, and file-sharing platforms.

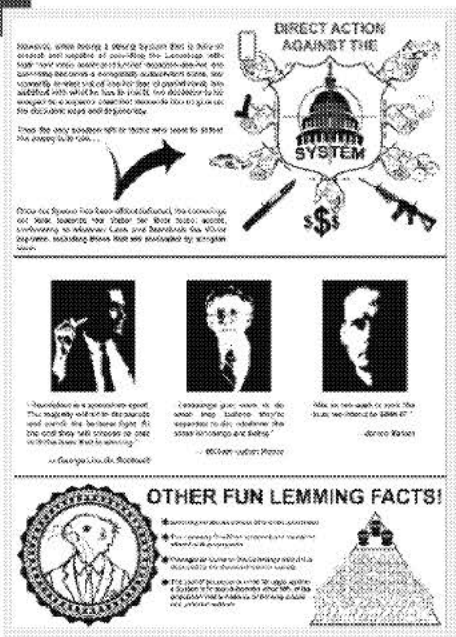
As reported by Tech Against Terrorism, the manifesto was originally uploaded on six file-sharing platforms and linked on **Twitter**. In the manifesto, the author stated that they would deliberately **target file-sharing platforms**: 'in case of early failure of my operation, these words get out to people'. It is highly likely that they believed hosting the manifesto on file-sharing platforms (such as Filemail.com, Zippyshare, files.safe.waifuhunter.club, delegao.moe, Mediafire and Anonfiles) would **ensure longevity** and discoverability of the content.

65 Image from Council for Media Services and Reset, *The Bratislava Shooting. Report on the role of online platforms*, p. 7 [LINK](#)

The Bratislava manifesto: borderline content and memes

The manifesto sets out the author's goals, motivations, inspirations, attack planning, and personal identity. The document specifically cites the perpetrators of several far-right terrorist attacks as inspiration.

66



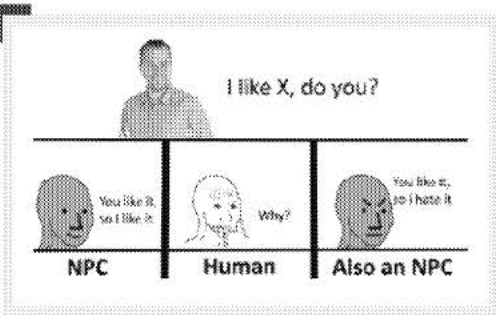
Antisemitic content shared by the attacker on page 9 of his manifesto, under the section titled ACCELERATE refers to key coded language such as ZOG, lemming theory: 'The only way to win, is to tear down this **rotten System**. Tear down the **ZOG**. Then, offer an alternative for the people, and those with brains will jump the **ZOG** ship before it sinks into the cold depths. First, you must understand the **lemming theory**. No one explains it better than the man who created it, William Luther Pierce.' And calls for violence against Jewish people: 'Whatever you dedicated to destroying ZOG. Not to attract the approval of average people – it doesn't matter. What you must do, is weaken **ZOG's hold on the populace**, by any means necessary.'

On page 37 of his manifesto he refers to the so-called 'reality pill' claiming 'Take the reality pill. Instead of **blackpills** and **whitepills**, focus on reality. The blackpill is thinking that all is doomed; that ZOG has already won; [...] The blackpill is thinking that our Race will never see victory. The whitepill is thinking

that we personally, will see victory. The reality pill is realizing that **our children will see victory**. We, the people alive at this moment, will not live to see the fruits of our work.'

On page 43 of the manifesto, the perpetrator refers to Russian policy and its far-right scene and how Russia, according to him, managed to pass from Bolshevism to being a 'White paradise'.

67



And he goes on referring to multiple well-known **antisemitic myths** claiming Jewish people are responsible for **mass migration** into Europe, for controlling governments, introducing the **Covid-19** vaccine, promoting **LGBTIQ** rights, **pornography**, for controlling the media and the entertainment industry, **social media** networks (the document includes references to Instagram, Twitter and TikTok specifically), the financial system, and for allegedly spreading ideologies such as Marxism and Progressivism.

The Slovakian Council for Media Services provided some examples of content glorifying the attack, which is still available in some cases, on major online platforms. The Council reports about comments made by TikTok users on news stories related to the attack. These comments praised the attack, albeit implicitly, and showed support for the attacker through the use of specific signs, which would only be understood by those who are native to 4chan or 8chan.

66 Image taken from the manifesto of the perpetrator of the Bratislava terrorist attack.

67 Image taken from the manifesto of the perpetrator of the Bratislava terrorist attack.

Examples of comments:

W strelec (w the shooter)

W: which on these platforms means 'win/winner'

Sigma: a pseudo-scientific construct of the hyper masculine alt-right denoting a particular type of male behaviour (successful, highly independent, intelligent). In effect, 'sigmas' are an equivalent of 'lone wolves'. As reported by the Slovakian Council, currently, the hyper masculine alt-right, best represented by the likes of Andrew Tate, is obsessed with the personality of Patrick Bateman from the movie *American Psycho* which, according to the community, best describes the qualities of 'sigmas'.

In both cases, the content becomes **illegal only after a careful analysis** and assessment (considering that one has the knowledge to do so).

The Slovakian Council for Media Services confirmed how **borderline content featuring emojis and slang** represents a formidable challenge for both regulators and social media platforms. As noted in a report by the Council released in cooperation with *Reset*, pre-emptive content moderation systems are simply unable to detect, and consequently remove, such content which significantly contributes to the spread of harmful/illegal content online.⁶⁸

Anti-LGBTIQ and anti-feminist content

Analysis and research carried out by think tanks and NGOs on violent extremists targeting women and the LGBTIQ community often reveal serious concerns about violence glorified and called for within online incel communities, revealing an increase in online references to inflicting violence and using degrading language against women on dedicated incel forums. Moreover, manifestos posted online by perpetrators of the Buffalo and Bratislava attack contained misogynist language and messages against women and feminism, as well as against LGBTIQ people.

Member States have underlined the presence of incel symbols in violent extremist content, including a 'chad' (alpha male) portraying Breivik (see image in the following pages). The Radicalisation Awareness Network (RAN Practitioners) has also organised events and provided material to support practitioners working on PCVE on how to deal with violent extremists within the incel community.⁶⁹ Since 2014, multiple public violent incidents (i.e., mass-shootings) have been tied to incels in the United States and Canada. Elliot Rodger, a 22 year old man, murdered six people and injured fourteen more in Isla Vista, California, before killing himself. Before he died, he posted a long manifesto and several videos on YouTube detailing his hateful worldview and membership to a nascent community of incels. Since then, dozens more victims have been murdered by self-proclaimed incels around the world.⁷⁰

In 2019 the perpetrator of the extremist right wing terrorist attack in Halle, Germany, on a synagogue defined himself as incel and blamed low birth rates in the West on feminism. He also linked this phenomenon to mass immigration, a clear reference to the 'Great Replacement' conspiracy theory. Likewise, the manifesto released by the perpetrator of the Hanau terrorist attack is riddled with conspiracist thinking and misogyny, with several pages titled 'Topic Women'. The attacker highlights his frustration with women, stating he had not been in a relationship for the last 18 years.

68 Council for Media Services and Reset, *The Bratislava Shooting. Report on the role of online platforms* [Link](#)

69 Video released by the RAN in 2021 <https://www.youtube.com/watch?v=sXIGZ7-258k>, conclusion paper by RAN C&N 'The incel phenomenon' on 28 July 2021 https://home-affairs.ec.europa.eu/system/files/2021-08/ran_cn_incel_phenomenon_20210803_en.pdf

70 *The Incelosphere. Exposing pathways into incel communities and the harms they pose to women and children*, Center for Countering Digital Hate, 2022.

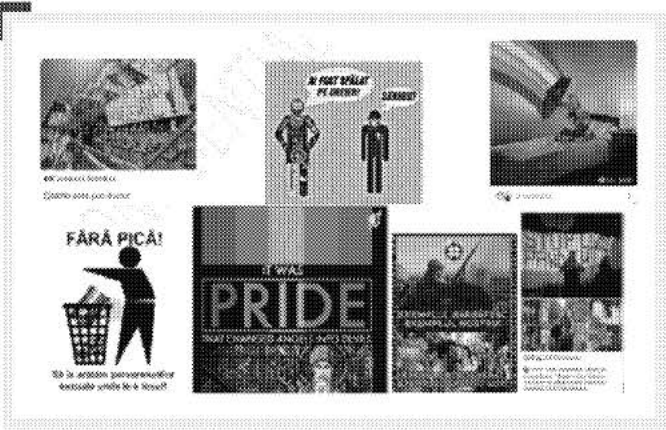
The outcomes of the RAN Communication and Narratives (C&N) Working Group meeting held in June 2021 on this subject revealed that the incel community is mainly present online on dedicated website, such as incels.is, now incels.co, as well as on mainstream social media like YouTube, Facebook and Twitter and less regulated platforms like 4chan and 8kun. Several incel forums and pages were banned, including a number of sub-Reddits, due to their harmful content. Incel forums are frequented throughout the EU, especially in Germany and country-specific themes can be found in the narratives they use, such as the role of height and race.

The RAN working group identified three levels of violence within incels to which relevant keywords can be associated:

- **Personal violence:** self-harm and suicidal ideation are common amongst incels. For example, the term **LDAR**, or '**Lay Down And Rot**', is frequently used on incel forums. Moreover, getting self-help or seeking mental health support are discouraged amongst incels.
- **Interpersonal violence:** incels are frequently encouraged to take others with them if they are going to harm themselves, i.e. that if they commit an attack, this should also harm other individuals. Moreover, women are harassed by incels both online and offline. An example is '**Chadfishing**', where an incel poses as a 'Chad' (i.e. the stereotypical handsome **alpha male**, according to incels) to get a date with a woman, but their true intention is to scorn and mock the woman. The 'Chads' they pose as are also targeted by putting down harmful texts under their fake profile (i.e. 'I'm a convicted **child molester**'). This way, incels can prove that it is indeed only the looks that matter for a woman that would still want to date such a person.
- **Societal violence:** violence aimed at society through mass shootings, for example intended for attractive women. Within the incel community, these mass shooters are often idolised and seen as examples. Moreover, elements of **gamification** can be found on incel forums (i.e. scoreboard ranking of perpetrators).

The RAN looked into the incel phenomenon from a Prevention and countering violent extremism (PCVE) perspective. It was confirmed that there are links between certain parts of the incel movement to (other) types of extremism, therefore it is important to try to understand how these people are interacting online.

71

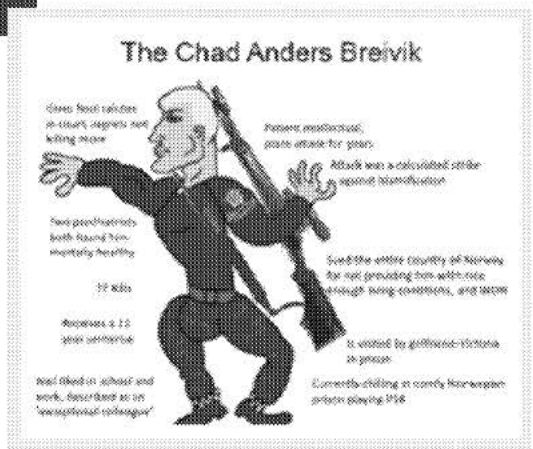


Anti-LGBTIQ content posted by violent extremists was observed by Member States. In Romania, this type of content is usually created by ultras groups⁷² and ultranationalists (individuals or organisations) and can be found not only on fringe platforms, but also on major ones like Facebook and Instagram. These types of messages are posted all year round, but see spikes in activity during June (**Pride Month**). Here are some visual examples of anti-LGBTIQ and **misogynist** content.

71 Images provided by Romanian authorities for the establishment of this handbook.

72 Ultras are a type of association football fans who are renowned for their fanatical support. The term originated in Italy, but is used worldwide to describe predominantly organised fans of association football teams.

Member States provided information about the so-called **Chad-Meme**, which stems from the incel subculture and is in some cases used in right wing extremist milieus to glorify perpetrators of terrorist attacks (see **Breivik's chad-meme** in the image). The meme represents an exaggeration of this masculine figure. The character depicted is called Chad. His disproportionate body strikes a humorous pose and gestures awkwardly. In Breivik's example, the image is integrated by explanation of the reasons why he should be considered as a winner.



Merchandise with slogans containing the term **'Groomer'** should be analysed closely to assess a link to **discriminatory narratives against LGBTIQ** and trans practitioners working with children. The term is used to justify hate, discrimination and violence against the LGBTIQ community by claiming that members of the community pose a danger to children. It implies that contact with children can lead to these children 'becoming queer'. The term is also used in the context of comparing homosexuality to child abuse. Its uses can include:

- 'OK Groomer'
- 'Hey, Groomer! Leave those kids alone!'
- 'Stop giving groomers access to your children'

The term **'roastie'** is slang used by the incel community to target and disparage sexually active women.



'Red pilling' refers to the so-called realisation that men do not have power or privilege in society. Contrary to this, men are vulnerable to being exploited by women in social, economic and sexual terms.

73 Image provided by German authorities for the establishment of this handbook

74 Source for the images: T. Squitoli, C. Martiny, *Profiting from hate*, ISD, 2022, pp. 27-31

Use of borderline and violent content within online incel communities

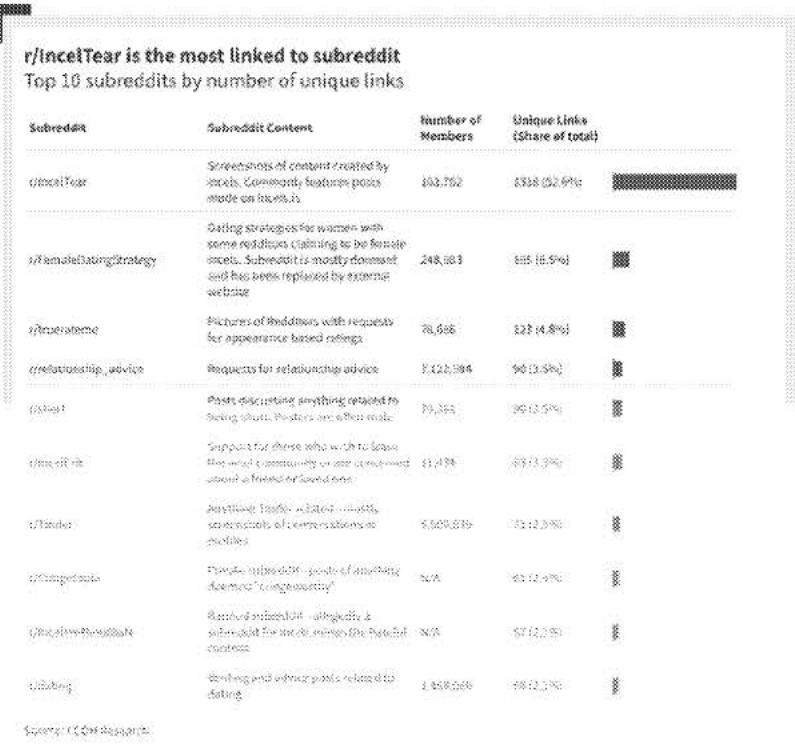
Researchers⁷⁵ from the Center for Countering Digital Hate released a report in December 2022 called *The Incelosphere. Exposing pathways into incel communities and the harms they pose to women and children*, which is based on the collection of all posts from the most popular subforum of the incel forum incels.co, posted from 1 January 2021 to 7 July 2022, leading to the creation of a dataset of 1,183,812 posts. The outcomes of this work revealed an alarming **high presence of harmful misogynist content** and a strong **link between incels and violent extremism**.

Authors of *The Incelosphere* report of a man being charged in July 2021 with illegal possession of a modified AR-15 type assault rifle with a suspected **'bump stock'** attachment that would allow for rapid fire, and one count of attempting to commit a hate crime relating to a plot to 'slaughter' women at an unnamed Ohio university. Whilst media outlets stated only that the perpetrator 'had been active on incel websites', members of the website under study suggested that he had posted on their site under the alias **'Oedipus'**.

The dataset reveals that the incel forum consists of a small number of active members with interest from a much larger number of visiting users. So-called **power users** (previously described as 'super spreaders') in the incel forum share an average of 23 posts a day and spend around 8 hours a day online.

Forum posts reveal promotion of extreme hatred, rape, pedophilia and mass shootings:

- Over a fifth of posts in the forum **feature misogynist, racist, antisemitic or anti-LGBTIQ language**, with 16% of posts featuring misogynist slurs.
- Forum members post about rape every 29 minutes, and examination of discussions of rape shows that 89% of posters are supportive.
- Posts mentioning **incel mass murders increased by 59%** between 2021 and 2022
- Analysis of the 'tags' applied to discussion threads on the forum shows that over a third are tagged with topics promoting expressions of anger or despair over members' incel status. Just 5.8% are tagged with topics promoting a more optimistic outlook.⁷⁶

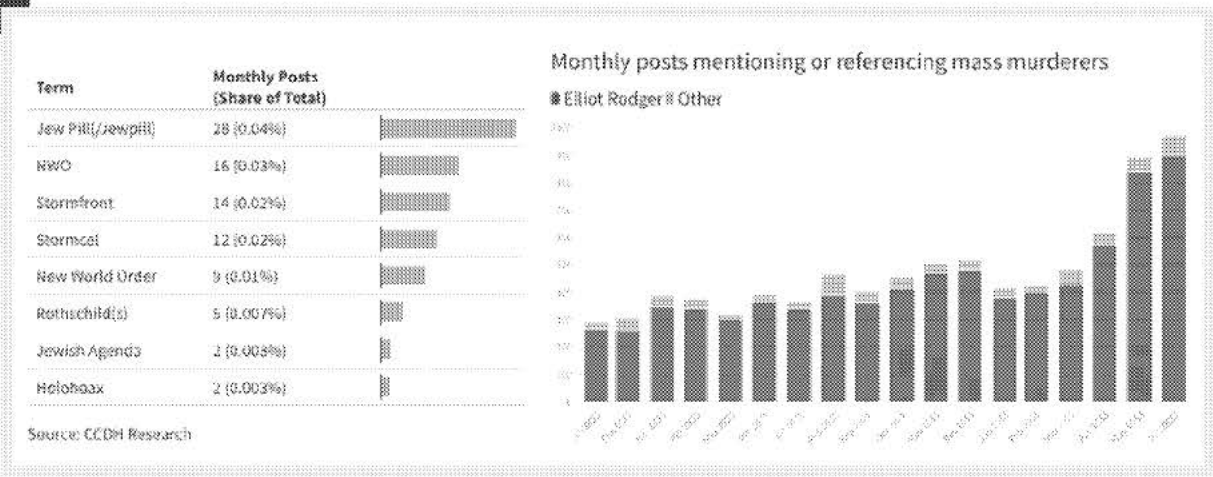
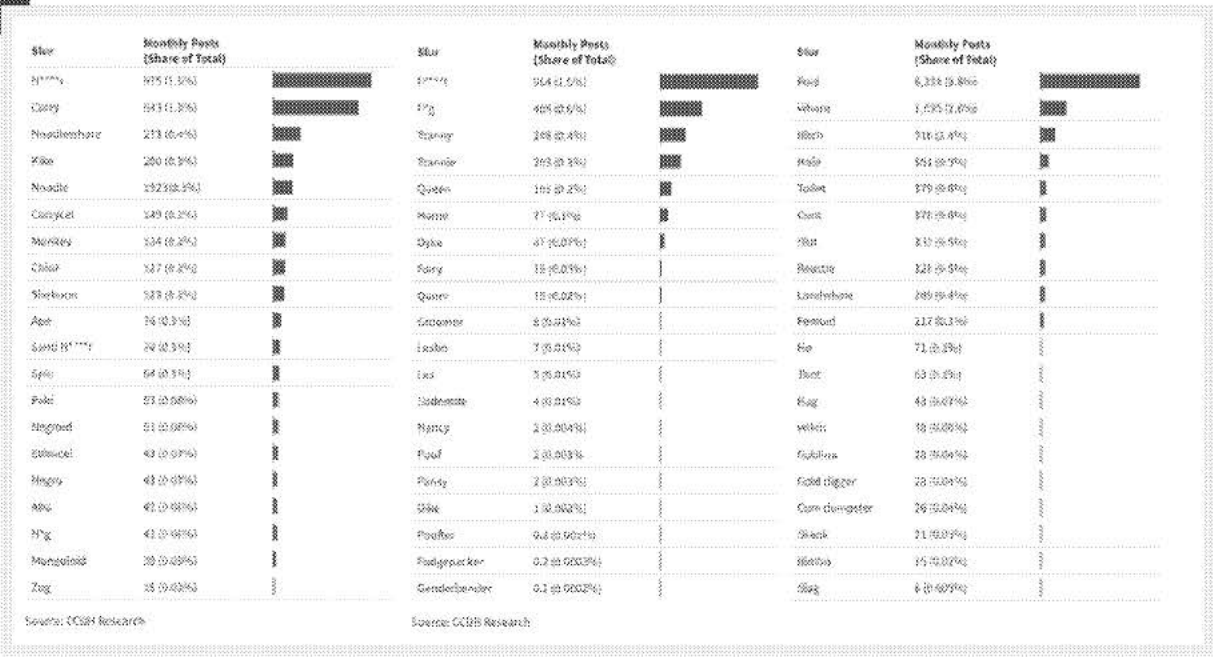


Most linked **sub-reddits** were also identified in the report (see table below). The report also refers to a channel called **'Sluthate Creeps'** which shows filmed videos of women and inter-racial couples in London with **misogynist and racist comments**. Comments to such videos can read: **'Blackcels I offer you beautiful white woman'**, or 'sometimes I do witness black ascension... I took footage'. This refers to the commonly held belief on the forum that **men from non-white backgrounds must 'date-down'**. Another identified title reads: 'The white race is the most beautiful' with thousands of views.

75 Center for Countering Digital Hate, *The Incelosphere. Exposing pathways into incel communities and the harms they pose to women and children*, 2022.
76 Ibidem, page 18

Misogynist, antisemitic and racist slurs encountered in the most popular incels forum are listed in the tables below.

77



77 Center for Countering Digital Hate, *The Incelsphere: Exposing pathways into incel communities and the harms they pose to women and children*, 2022.

Policy perspective: what are the current measures in place to prevent the spread of borderline content

Legislative and voluntary approaches in the EU

The EU has established a series of legislative and non-legislative actions to address the threat posed by the use and spread of disinformation/misinformation and hate speech for malicious purposes.

Voluntary agreements, codes of conduct and practices, as well as legislative measures were taken in the EU to reinforce the cooperation between governments, civil society and online companies to curb the spread of illegal and harmful content. Hereby we will focus on the actions taken at EU level to prevent the dissemination of harmful but lawful content that can lead to radicalisation and violent extremism.

The voluntary/cooperative approach

The EU Internet Forum

The EU Internet Forum (EUIF) launched by the Commission in December 2015, addresses the misuse of the internet for terrorist purposes through two main strands of action:

- Reducing accessibility to terrorist content online
- Increasing the volume of effective alternative narratives online

Enhancing the fight against child sexual abuse online, was added to EUIF's area of activities in 2019.

The EUIF is instrumental in pushing forward a number of initiatives addressing the European Union and industry response to terrorist attacks with an online dimension. On content moderation, the EUIF has taken a pragmatic approach in providing support to the industry on how to best identify content referring to fragmented, hard-to-define ideologies and legal but harmful content that can lead towards radicalisation. Among the actions taken so far in this context are:

- The EUIF involvement in the creation of Europol's EU Internet Referral Unit (EU IRU) which flags and refers terrorist content online to online platforms.
- The establishment of the EU Crisis Protocol to respond to the viral spread of terrorist and violent extremist content online.
- The development of a list of violent right wing extremist groups, symbols, and manifestoes aimed at facilitating online content moderation for industry stakeholders to respond to the challenges posed by Violent Right-Wing Extremism's (VRWE) presence online.

The EU Internet Forum has committed since 2020 to look into the effects of algorithmic amplification of TVEC and borderline content on the user journey towards radicalisation.

Actions taken:

- Two workshops and a study to understand the potential negative effects of the use of algorithmic amplification to spread TVEC and borderline content
- Guidance to tech companies on identification of borderline content and tactics used to spread it through this handbook
- RAN Stratcom workshop to find Stratcom solution to the risks posed by the misuse of new tech and algorithmic systems
- Cooperation with international partners
 - Co-lead of Christchurch Call workstream on algorithmic amplification and positive interventions
 - Members of GIFCT working group on technical approaches, which also addressed the impact of algorithmic amplification spreading borderline content

The initiative of establishing this handbook originates from the outcomes of the two workshops held in 2021 and 2022 on the subject, as the automated amplification of borderline content has been identified as particularly problematic, due to the lack of clear guidance on thresholds and definitions, which makes it hard for online platforms to moderate this content and prevent it from being recommended.

The legislative approach

The Eu Digital Services Act

The Digital Services Act (DSA)⁷⁸ is a *horizontal instrument* applicable to providers of intermediary services offered in the European Union. Seeking to create a safer digital space for all, the DSA sets rules on the content moderation practices of online platforms, in particular the removal of illegal content, and their interaction with freedom of speech. It also creates a stronger public oversight of online platforms, in particular for platforms that reach more than 10% of the EU's population.

Which providers are covered?

Under the DSA, the obligations of different online players match their role, size and impact in the online ecosystem:

- **Intermediary services** offering network infrastructure: internet access providers, domain name registrars, internet service providers, cloud services, messaging, marketplaces, or social networks.
- **Hosting services** such as cloud and web hosting services, including also:
 - **Online platforms** bringing together sellers and consumers such as online marketplaces, app stores, collab economy platforms and social media platforms.
 - **Very large online platforms** (VLOPs and very large online search engines (VLOSEs) pose particular risks in the dissemination of illegal content and societal harms. Specific rules are foreseen for those VLOPs and VLOSEs reaching more than 10% of 450 million consumers in Europe.

All online intermediaries offering their services in the single market, whether they are established in the EU or outside, will have to comply with the new rules. Micro and small companies will have obligations proportionate to their ability and size while ensuring they remain accountable. In addition, even if micro and small companies grow significantly, they would benefit from a targeted exemption from a set of obligations during a transitional 12-month period.

⁷⁸ [Link to the European Commission's official page](#) [LINK](#)

How will citizens benefit from the DSA?

The DSA will create a safer online experience for citizens to freely express their ideas, communicate and shop online, by reducing their exposure to illegal activities and dangerous goods and ensuring the protection of fundamental rights.

Some of the DSA obligations relevant for the scope of this handbook include:

- **Effective safeguards for users**, including the possibility to challenge platforms' content moderation decisions based on new obligatory information to users when their content gets removed or restricted.
- **Wide ranging transparency measures for online platforms**, including better information on terms and conditions, as well as publicly available transparency reports on the algorithms used for recommending content or products to users.
- New obligations for the **protection of minors** on any platform in the EU.
- Obligations for very large online platforms and search engines to **prevent abuse of their systems by taking risk-based action**, including oversight through independent audits of their risk management measures. Platforms must mitigate against risks such as **disinformation** or election **manipulation**, cyber violence against women, or harms to minors online. These measures must be carefully balanced against restrictions of freedom of expression, and are subject to independent audits.
- New provisions to **allow access to data to researchers** of key platforms, in order to scrutinise how platforms work and how online risks evolve;
- The **liability rules for intermediaries have been reconfirmed** and updated by the co-legislator, including a Europe-wide prohibition of generalised monitoring obligations.

The code of practice against disinformation

Major online platforms, emerging and specialised platforms, players in the advertising industry, fact-checkers, research and civil society organisations delivered a strengthened Code of Practice on Disinformation following the Commission's Guidance of May 2021.

The **strengthened Code of Practice on Disinformation**⁷⁹ has been signed and presented on the **16 June 2022** by 34 signatories who have joined the revision process of the 2018 Code.

The new Code aims to achieve the objectives of the Commission's Guidance presented in May 2021, by setting a broader range of commitments and measures to counter online disinformation.

The 2022 Code of Practice is the result of the work carried out by the signatories. It is for the signatories to decide which commitments they sign up to and it is their responsibility to ensure the effectiveness of their commitments' implementation. The Code is not endorsed by the Commission, while **the Commission set out its expectations** in the Guidance and considers that, as a whole, the Code fulfils these expectations.

Signatories committed to take action in several domains, such as; **demonetising** the dissemination of disinformation; ensuring the **transparency** of political advertising; **empowering** users; enhancing the cooperation with **fact-checkers**; and providing researchers with better **access to data**.

⁷⁹ [Link to the European Commission's official website LINK](#)

The code of conduct on countering illegal hate speech online

The Framework decision (2008/913/JHA)⁸⁰ of 28 November 2008 on combating certain forms and expressions of racism and xenophobia is the legislation addressing illegal hate speech in the EU.

To prevent and counter the spread of illegal hate speech online, in May 2016, the Commission agreed with Facebook, Microsoft, Twitter and YouTube a 'Code of conduct on countering illegal hate speech online', as defined by the Framework Decision.

Throughout 2018, Instagram, Snapchat and Dailymotion took part to the Code of Conduct, Jeuxvideo.com in January 2019, TikTok in 2020 and LinkedIn in 2021. In May and June 2022, respectively, Rakuten, Viber and Twitch announced their participation to the Code of Conduct. The implementation of the Code of Conduct is evaluated through a regular monitoring exercise set up in collaboration with a network of organisations located in the different EU countries. A common methodology is used to test how the signatories are implementing the commitments.

Under the Code of Conduct on hate speech⁸¹, the signatory IT companies committed to the following activities which are relevant in light of disrupting economic activities that represent or promote dangerously hateful and extremist ideologies or actors:

- Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct.
- Intensify cooperation between themselves and other platforms and social media companies to enhance best practice sharing.

The European Observatory of Online Hate Speech

The European Observatory of Online Hate⁸² is one of the most recent projects addressing online hate, supported by the European Commission's Rights, Equality and Citizenship Programme awarded by the Commission (DG Justice). The project is being implemented by a consortium of 4 partners with a proven track record of successful implementation and impact. The project is being brought to life with *Textgain* in the lead, and in cooperation with *Dare To Be Grey*, *Hogeschool Utrecht* and *PCDS*.

The consortium is tasked with conducting a **two-year investigation into and reporting on the fundamental nature of the dynamics of online hate**, how hate manifests itself, the connections between the perpetrators and their influence as well as disinformation strategies. This investigation will involve the collaborative development of a monitoring tool using cutting edge AI tools developed by Textgain.⁸³ The observatory is monitoring 15 social media platforms and providing early detection of hate speech/disinformation for all 24 European languages (+ Arabic, Russian, Turkish and more).

80 Link to the European Commission's official website [LINK](#)

81 More information on the European Commission's website: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

82 More information on EOOH's official website: <https://eoooh.eu/about-us>

83 More information on Text Again website: <https://www.textgain.com/>

Applying definitions of borderline content in relation to violent extremism into policy guidelines

The EUIF selected examples of **EUIF tech companies' guidelines and terms of services**, of unacceptable harmful content and disinformation. Such efforts aim at preventing the spread not only of illegal content, but also of 'harmful but lawful' content that can lead to violent behaviour. For the scope of this handbook, we have selected the content addressed in the guidelines that can lead to radicalisation and violent extremist behaviour. By providing these examples, the European Commission (EUIF) does not intend to endorse these companies' content moderation practices, but rather to **share examples of best practices in defining of borderline content** leading to violent extremism and inform about companies' activities on that front. Further reference and analysis of tech companies' internal existing policies is also provided by GIFCT, in relation to GIFCT member platforms, in Annex.

Discord's advanced community management

Discord guidance to content moderation includes reference to most common hateful slurs and expressions that should be moderated by the platform and it is divided into two groups:

1. Racist Terms

- **Jap:** Used during World War II when the Japanese were in interment camps in the US, this term was a derogatory way US citizens referred to Japanese people and is heavily considered an ethnic slur against Japanese people;
- **Gypsy/Gypped:** Both to refer to 'Gypsy' the people, and to be robbed/conned in the form of 'gypped', this is a term specifically used as an ethnic slur against the Romani people. While it is used in legal contexts, the words have slowly been brought out of use due to its common use as a slur historically.
- **Chink/Ching Chong:** Chink has been historically used as a slur against people of Chinese descent, and sometimes even Asian decent widely, with ching chong mocking the language of the Chinese which is commonly used alongside chink.
- **Triple Parentheses, also known as (((echo)))**: This is a very uncommon but recently used symbol to denote someone of Jewish origin, typically in a way to target or harass them. This symbol is used to single Jewish people out by communities and places a target on their back for their religion or ethnicity, and should not be tolerated.

2. LGBTIQ Specific Slurs

- **Dyke/Lesbo:** A term originated as a slur against more masculine-presenting lesbian women, this term has been reappropriated by its community into being a common slang term to refer to lesbian women. While some people would not mind being called a dyke, it should be made aware of its possible negative downsides for people who may be uncomfortable with the term.
- **Thing:** Specifically in reference to pronouns, the use of 'thing' instead of a user's preferred pronouns, used to usually mock a user's preferred way of expressing their gender identity, is commonly used as a way to invalidate or minimise trans/enby people.
- **Fag/Faggot/Homo:** All terms used to refer to gay people, and all heavy slurs with the intention of belittling and attacking them. These words are also commonly used in the real world when they are attacked, and should not be taken lightly.

- **Trap:** A term that originated from anime, this word is in reference to men who dress as females and look female-presenting, and 'trap' heterosexual people into having an attraction to them. This word has been used out of its original contexts as a slur to transgender people, as if their existence is to 'trap' or 'trick' people around them. Not everyone finds this term offensive. It should be evaluated as to whether it deserves moderation on a case-by-case basis.

While some of these terms may be popular in certain spaces (such as gaming), it is important to understand the history and weight behind them, and act accordingly about their place in long-term server.

Meta

Meta does not accept any type of degrading and **dehumanising speech** in its platform. The platform defines attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. Moreover, any type of harmful stereotypes, which Meta defines as dehumanising comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence, are prohibited.

Meta also prohibits the usage of **slurs** that are intended to attack people on the basis of their **characteristics**. However, the social media recognise that people sometimes share content that includes slurs or someone else's hate speech to condemn it or raise awareness. In other cases, speech, including slurs that might otherwise violate their standards can be used self-referentially or in an empowering way. Their policies are designed to allow room for these types of speech, but do require people to **clearly indicate their intent**. If the intention is unclear, the content may be removed.

Furthermore, Meta has a policy on not allowing posts that target a person or group of people (including all groups except those who are considered non-protected groups described as having carried out violent crimes or sexual offenses or representing less than half of a group) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form.
- Dehumanising speech or imagery in the form of comparisons, generalizations, or unqualified behavioural statements (in written or visual form) to or about.
- Animals that are culturally perceived as intellectually or physically inferior or filth, bacteria, disease and feces.
- Sexual predator or other criminals (including but not limited to 'thieves,' 'bank robbers,' or saying 'All [protected characteristic or quasi-protected characteristic] are 'criminals').
- Statements denying existence.
- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image.

Designated dehumanising comparisons, generalisations, or behavioural statements (in written or visual form) that include:

- Black people and apes or ape-like creatures or farm equipment.
- Caricatures of Black people in the form of blackface.
- Jewish people and rats and Jewish people running the world or controlling major institutions such as media networks, the economy or the government.
- Denying or distorting information about the Holocaust.
- Muslim people and pigs or Muslim person and sexual relations with goats or pigs.
- Women as household objects or referring to women as property or 'objects'.
- Transgender or non-binary people referred to as 'it'.
- Dalits, scheduled caste or 'lower caste' people as menial labourers.

Other statements of inferiority, which are defined as:

- Expressions about being less than adequate, including but not limited to: worthless, useless.
- Expressions about being better/worse than another protected characteristic, including but not limited to: 'I believe that males are superior to females'.
- Expressions about deviating from the norm, including but not limited to: freaks, abnormal.
- Expressions of contempt (in written or visual form), which we define as: self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist.

Cursing, except certain gender-based cursing in a romantic break-up context, defined as:

- Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole.
- Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker.
- Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit.

Content that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels for the above characteristics.

Misinformation

One of the main targets of Meta to tackle misinformation is spotting fake accounts. Part of Meta's strategy is also to investigate and take down covert foreign and domestic influence operations that rely on fake accounts. Over the past three years, Meta has removed over **100 networks of coordinated inauthentic behaviour (CIB)** from their platform and kept the public informed about their efforts through on their monthly CIB reports. Meta has also been cracking down on deceptive behaviour and found that one of the best practices to fight this behaviour is by **disrupting the economic incentives** structure behind it. In that regard, the company have built teams and systems to detect and enforce against inauthentic behaviour tactics behind a lot of clickbait. Part of their strategy is also using artificial intelligence to help them detect fraud and enforce policies against inauthentic spam accounts.

Since misinformation can also be shared by people in good faith, Meta have built a global network of more than 80 **independent fact-checkers**, who review content in more than 60 languages. When a content reviewer rates something as false, the distribution is **automatic reduced and consequently fewer people see it**. Likewise, it adds a warning label with more information for anyone who sees it. Moreover, the platform notifies the person who posted it and reduces the distribution of pages, groups, and domains who repeatedly share misinformation.

When the disinformation is topics like COVID-19 and vaccines and content that is intended to suppress voting, the content is immediately removed. On Facebook and Instagram, the company started showing educational pop-ups connecting people to information from official sources linked to the topic they wanted to address – for example with COVID, Russian's invasion of Ukraine, or even in some elections in countries such as Brazil or Kenya.

Zoom community guidelines

To ensure that content considered abusive is removed, Zoom Trust & Safety team includes lawyers and engineers, as well as data science, security, privacy, product, and various other technical experts. When a report of abuse is sent, the team receive and evaluate the violations reported from the client, and remove it if considered abusive.

Zoom defines hateful conduct as a behaviour that promotes violence against or directly attacks or threatens other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease, prohibiting any promotion of hateful conducts. Equally, users are not allowed to use their user name, display name or profile information to abuse or threaten anyone. Accounts that do so may be permanently suspended. Moreover, organisations that promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease are not allowed on the platform.

Moreover, Zoom does not allow hateful imagery, which includes logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

The platform qualifies hateful imagery the following:

- Symbols historically associated with hate groups (e.g., the Nazi swastika).
- Images depicting others as less than human, or altered to include hateful symbols (e.g., altering images of individuals to include animalistic features).
- Images altered to include hateful symbols or references to a mass murder that targeted a protected category (e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust).

Youtube

YouTube does not allow content that encourages dangerous or illegal activities that risk serious physical harm or death, nor content that intend to praise, promote, or aid violent criminal organisations. These organisations are not allowed to use YouTube for any purpose, including recruitment. Likewise, content promoting violence or hatred content that promotes violence against individuals or groups based on any of the following attributes, is not allowed and will be removed:

- Caste
- Disability
- Ethnicity
- Gender Identity and Expression
- Nationality
- Race
- Immigration Status
- Religion
- Sex/Gender
- Sexual Orientation
- Victims of a major violent event and their kin
- Veteran Status

Suggested guidelines

In the following sections, inputs received by Member States and Civil Society to support actions by tech companies and other stakeholders in curbing the spread of borderline content are presented. This guidance is a collection of inputs received by EUIF stakeholders.

Definitions, taxonomies and content moderation

Focus on tactics and behaviour

Common best practices recommended relate to the need of **algorithmic transparency** and, more in general to the need to focus not only on content, but also on manipulation tactics used to amplify its spread and to manipulate users.

Recent consultations with EUIF members led to the following considerations on how to best develop actions and measures.

Both **safety by design measures** and **positive interventions** need to be adopted in order to prevent the spread of borderline content via automated systems.

As highlighted throughout this handbook and in the meetings held so far by the EU Internet Forum, governments and the industry's focus on content is not enough and can bring about major challenges that could be partly overcome through the establishment of common definitions, better complaint and redress mechanisms in case of removals and the need to preserve the fundamental right to freedom of speech.

Therefore, preventive measures should take into account **behavioural patterns and propagation tactics** used by malicious actors, including **manipulation**, and include digital and media literacy, critical thinking and democracy-strengthening programs to foster resilience.

Taxonomies

EUIF members also agreed that an attempt should be made to reach an agreement on some **definitions** and **taxonomies** on borderline content leading to violent extremism, which are essential basics to build effective measures.

Civil Society Organisations should be empowered to compile lists of **slurs, symbols, images, memes** (and other forms of harmful online material) **for content moderation** that complement and build on the efforts of national authorities.

This handbook is a first attempt to provide the basis for the creation of such taxonomies thanks to the support of GIFCT and other civil society organisations involved.

Avoid content moderation circumvention

In its recommendation developed for tech platforms in *Mitigating Content Moderation Circumvention*⁸⁴ on its Knowledge Sharing Platform, Tech Against Terrorism raised a point on nuances in extremist content. Therefore, the NGO advises to:

- **'Improve human reviewers' understanding of trends** in terrorist and violent extremist use of the internet, including communication. Specialist knowledge is required to account for the nuances in extremist content and reduce the likelihood of false positives and false negatives. Contextual nuances such as cultural and political specificity, coded language, humour, satire, and irony are currently better judged by humans with specialist knowledge on a topic, language, cultural, or political situation. This includes understanding new and emerging violent extremist groups, and of local context.'
- 'Introduce a time limit on the validity of join links for private and encrypted servers and channels, to allow users to benefit from them whilst reducing terrorist and violent extremist actors' capacity to easily store and share them. Though the use of private and encrypted services should be upheld, platforms can strengthen **metadata and behavioural analysis of unencrypted metadata**, including names and photos of groups, geographic location, and member traffic.'
- 'Develop **existing content moderation practices beyond word and hashtag bans**, and invest in advanced automated content moderation tools to better identify lexical variations and deliberate misspellings of banned keywords and phrases. This includes broadening the scope of words and phrases banned, as well as including identifying commonly replaced letters for numbers, common slight misspellings, and using plural versions of words. However, Tech Against Terrorism recommends proceeding with caution when broadening the scope of banned key words and phrases so as to not compromise freedom of expression.'
- As regards the use that tech companies can make of the Knowledge Sharing Platform provided by TAT, the recommendation is to **regularly update the keywords and imagery** listing to catch new indicators. Platforms can find TAT's list on the KSP.

84 The TAT full brief is available on the Tech Against Terrorism's Knowledge Sharing Platform [LINK](#)

Standards and support to tech platforms

GIFCT-linked Tools to Support Tech Companies in Countering Terrorist and Violent Extremist Content

There are a range of platforms, particularly smaller or lesser resourced platforms, that may encounter more challenges when implementing measures and standards to limit the spread of borderline content in relation to violent extremism. The following table established by GIFCT provides an overview of existing and developing tools and practices that tech platforms can use as support to their content moderation efforts.

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 ⁸⁵ to help other members understand what content corresponds to the hash, including content type, terrorist entity that produced the content, and its behavioural elements. A 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 AI models that developed over the past five years to classify Daesh and al-Qaeda propaganda in multiple formats with exceptionally high performance. ⁸⁶
Meta's Hash Matcher Actioner	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. ⁸⁷ Hasher-Matcher-Actioner (HMA) 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. ⁸⁸ HMA 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 Against Terrorism's moderation tool	Being developed by Google's research and development unit, Jigsaw, in partnership with Tech Against 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. ⁸⁹
Tech Against Terrorism's TCAP	The Terrorist Content Analytics Platform (TCAP) seeks to disrupt terrorist use of the internet by facilitating the quick and accurate removal of terrorist content. ⁹⁰ It does this by alerting terrorist content to tech companies when found on their platforms. A team tracking terrorist migration across a variety of tech platforms flag URLs containing terrorist content to the TCAP. TCAP sends alerts to tech platforms about terrorist content found on their sites and checks the status of URLs to determine when content is removed.

85 See the 2022 GIFCT Annual 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>

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

87 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/>

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

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

90 TCAP was developed by Tech Against Terrorism. For more information on the TCAP: <https://www.terrorismanalytics.org/>

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. Accordingly, 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.**

Demonetisation

Another important aspect that must be kept in mind when developing measures against the spread of harmful borderline content is the increasing use of platforms for economic and activities aimed at raising money for violent extremist groups. During the **EUIF workshop on Violent Extremist and Terrorist Financing Online** participants agreed on a series of recommendations to tech companies to:

- Develop, implement and enforce robust **terms of services (ToS) or Community Guidelines** prohibiting malicious actors and organisations, brands/labels, marketplaces or online stores that engage in activity with financial gains that promote, supports or glorify dangerous disinformation, hate speech, racist content, and gender-based violence and hate.
- Consider **off-platform behaviour** when assessing the user behaviour, including external web profiles of sellers to judge whether they are likely **funding hate groups and extremists**, and adopt measures to enable the **verification and assessment** of the landing/destination page of outlinks to online stores and crowdfunding websites, also using internal crawlers to **identify** those **outlinks**.
- **Partner with CSOs to be up-to-date on terms associated with hateful and extremist ideologies.** Consider, abbreviations and coded terms in automated and manual content moderation.
- Consider the **meaning of content in the wider context** to assess if it promotes dangerous hateful or violent extremist ideology or activity.
- Do not allow **ads** and sponsorships on search terms, content, users and activities that are related to dangerously hateful, extremist ideologies and conspiracy theories.
- Avoid actively **recommending products**, content, sites and users that fall in the above listed category.

Risk assessments

As highlighted by Member States in the EUIF workshop on borderline content in September 2022, several challenges encountered in the identification and definition of borderline content, and especially disinformation, are linked to the **assessment of the intention**. Regarding disinformation, for instance, it is in some cases difficult to distinguish between false information spread with malicious intent and misinformed opinions.

Challenges are also linked to finding the right **balance** between protecting users from malicious content leading to violence and the fundamental right to freedom of speech and expression.

Among the **strategic objectives** identified by Italy and other EU Member States⁹¹ as key actions to curb the spread of borderline content, especially disinformation, leading to violent extremism are:

- The importance of carrying out **risk assessments** of the content to verify if the information is false and intentionally misleading. The risk assessment process must be based on **specific indicators** concerning the assessment of the kind of information, the sources, the consequences of the dissemination of such content.
- The nature of the threat posed by harmful but lawful content that can radicalise users requires an approach at **international level** that must take into account the needed balance between promoting users' safety and preserving their fundamental right to freedom of speech.
- All stakeholders involved in the **EUIF** and other **international fora** must make full use of **tools** they have at disposal to face the threat that borderline content may pose to our democratic values and to users safety.
 - The **EU Internet Forum and the Global Internet Forum to Counter Terrorism** will continue to work together to establish such instruments and guarantee support to tech companies, especially small ones, also with the aid of Europol's Internet Referral Unit and Tech Against Terrorism.
- EU Member States should continue to refer to the **EU Internet Referral Unit** in Europol to ensure coordination and cooperation between each other and with tech companies affected by the dissemination of harmful content.

Member States are well aware of the challenges encountered by platforms in the moderation of borderline content, because of the need to safeguard fundamental rights, especially freedom of speech, and because of the difficulty in reaching agreements on common definitions.

91. These recommendations have not been officially endorsed by the European Commission.

Mitigation measures: the role of recommender systems

According to EU Member States, tech companies should focus their preventive measures on making the internet users, especially young ones, **more resilient** towards borderline content leading to violent extremism, not necessarily through removals, but also by limiting its spread and visibility.

In order to promote better media resilience among **young people**, which remain the most vulnerable group, the Netherlands suggested for online platforms to work more on the prevention of

- Active **recruitment** on alt tech platforms, which more often there appears to be an overlap of different themes, and the **rabbit hole effect**.
- **Addressing algorithmic amplification**, one of the most important enabling mechanisms of the massive disseminating of borderline content.

As suggested by EU Member States, tech companies should **focus** on those systems that enable the spread of borderline content, such as **recommender systems/content sharing algorithms** rather than on removing the (borderline) content itself.

It is yet still unclear how these algorithms work and their effect on the radicalisation process, due to a lack of transparency regarding the way they function. The Netherlands suggested the following points of focus for future work:

- **The Digital Services Act** will provide vetted researchers access to algorithms of VLOPS (Very Large Online Platforms). Therefore, it is very important to have enough **skilled researchers to unravel these algorithms** and to have a shared research agenda at the European level.
- **Transparency** of algorithms is key, but not the only measure to tackle algorithmic amplification – more measures are needed.
- Member States and the Commission can also focus on creating **technical interventions** (together with tech companies) which restrain the dissemination of borderline content:
 - Creating **(pause) buttons** on social media platforms that make it harder to copy and disseminate ('retweet') borderline content.
 - Adding more **warnings** to remind internet users of their internet policies regarding posting and creating extreme content.

Policy efforts need to be stepped up to tackle the role of algorithms in the radicalisation process.

The Czech Republic suggested it is important to continue the discussion at EU level on disinformation and propaganda, as well as the topic of algorithmic amplification, which facilitates the dissemination of this content. The Czech Republic recommends to **turn these discussions** into a **genuine effort** to draw practical consequences on the subject, building on existing initiatives such as: a strengthened Code of Practice on Disinformation or the EU ban on RT, Sputnik and several other media outlets.

The Czech Republic believes in a consensus on the seriousness of this issue and in the effort to jointly find an effective solution that would **strengthen the resilience and security** of European states against the influence operations of aggressors.

While there is a need to ensure that algorithmic amplification techniques are not used to spread TVEC and borderline content leading to radicalisation, multi-stakeholder fora have also looked into ways in which recommender systems can be used to promote the dissemination of alternative and positive narratives that can shift users' attention away from harmful content. In the context of counter narratives,

EUIF has established the Civil Society Empowerment Programme (CSEP)⁹² under which Civil Society Organisations are receiving funds to implement counter narrative campaigns. Through CSEP, the EU is committed to capacity building, training, partnering civil society organisations with internet and social media companies, and supporting campaigns designed to reach vulnerable individuals and those at risk of radicalisation and recruitment by extremists. Such instruments are crucial to avoid polarisation and build resilience among online users against malicious content.

The EU Fundamental Rights Agency on bias in algorithms

Algorithms can be used for **ranking content or to detect and flag hate speech**. When it comes to the use of algorithms to prevent the spread of offensive speech, a report released by the EU Fundamental Rights Agency (FRA) in 2022 confirms the very real danger of biased algorithms that may lead to discriminatory outcomes against protected groups. The report shows, on the one hand, that the development of bias in algorithms over time, through feedback loops, risks reinforcing or creating discriminatory practices that affect vulnerable groups disproportionately. Small differences and biases may amplify over time, if no mitigation actions are taken.

FRA also observed how offensive and **hate speech detection algorithms**, based on advanced machine learning methodologies and natural language processing (NLP), still strongly rely on certain terms and words and contain a variety of biases. Given the sheer magnitude of online content, major online platforms have considerably increased their efforts to automatically detect or 'predict' potential online hatred, and have developed tools to do this. However, **such tools can produce biased results** for several reasons.⁹³

Most notably, the level of **hatred** associated with different identity terms (i.e. words indicating group identities, such as Muslim, refugee or Jew) **varies** considerably **across the data and models** that form the basis of the tools. For example, sentences using the term 'Jew' in English-language models lead to a much greater increase in the predicted level of offensiveness than the term 'Christian'. This leads to differences with respect to the **predictions of offensive speech** for different groups. Those differences can also lead to wrong predictions and classifications.

For the **development of the report produced** by FRA, several algorithms for offensive speech detection were specifically developed, based on different methodologies and for different languages – English, German and Italian – and subsequently **tested for bias**. The outcomes show that some terms lead considerably more often to predictions of text as being offensive. For example, in English, the use of terms alluding to 'Muslim', 'gay' or 'Jew' often lead to predictions of generally **non-offensive text phrases as being offensive**. In the German-language algorithms developed for this report, the terms 'Muslim', 'foreigner' and 'Roma' most often lead to predictions of text as being offensive despite being non-offensive. In the Italian-language algorithms, the terms 'Muslims', 'Africans', 'Jews', 'foreigners', 'Roma' and 'Nigerians' trigger overly strong predictions in relation to offensiveness.

On the other hand, terms not usually linked to hate speech may be disproportionately often missed. People expressing **hate speech can easily avoid being detected through simple changing of words that may be more easily flagged**. One of the reasons for these results is that these terms are strongly linked to online hatred captured in the 'training data' (text datasets including examples of hatred) used for creating the algorithms.

Therefore, FRA recommends assessing **potential disproportionate 'overpolicing' of certain groups**, and to carry out assessments of outputs (algorithmic predictions) with respect to the composition of the target groups.⁹⁴

⁹² Reference link

https://home-affairs.ec.europa.eu/networks/radicalisation-awareness-network-ran/civil-society-empowerment-programme_en

⁹³ Ibidem, page 13

⁹⁴ EU Fundamental Rights Agency (FRA), Bias in Algorithms. Artificial Intelligence and Discrimination, Vienna 2022, page 12

The differences in bias in algorithms found in this project differ not only across languages but also across the tools and methodologies used. Some **AI models** are based on algorithms that were developed for general language detection and prediction tasks based on large bodies of text. Such '**pre-developed**' **models** are needed for speech detection algorithms in many cases. The research found that **bias differed depending on which tools were used**.

This means that bias is already embedded in **predeveloped general-purpose language models**, which are often developed by large companies with access to vast amounts of data and computing power. Assessing and documenting bias in such pre-developed models is challenging in the absence of full documentation and available tests for identifying bias in such tools. In addition, datasets are often difficult to obtain. This is partly because NLP researchers are overly cautious and avoid sharing data, often because they lack knowledge of data protection rules.⁹⁵

FRA highlights the need to **strive for more language diversity in NLP tools**, and calls upon the EU and its Member States to consider building a repository of data for bias testing in NLP. Such a repository, in conformity with data protection rules, could contain data in all EU languages to enable biases testing on a continuous basis.⁹⁶

In its **2020 report 'Getting the future right'**,⁹⁷ FRA also highlighted the need for further studies of potential discrimination resulting from the use of AI systems. Some professionals FRA interviewed for this report underscored that results from complex machine learning algorithms are often very difficult to understand and explain. This leads to the conclusion that further research to better understand and explain such results ('explainable AI') could also help to better detect discrimination when using AI.⁹⁸

The FRA recommends **that Article 40 of the EU Digital Services Act**, which allows for researchers to **access** data from very large online platforms and very large online search engines, be used to the extent possible, to allow **access to data needed for bias and discrimination** related research on online platforms' conduct.⁹⁹

FRA is continuing to work on understanding and analysing fundamental rights related risks of using algorithms and AI.¹⁰⁰ At the same time, FRA is investigating questions linked to online content moderation with respect to addressing hate speech in a fundamental rights compliant manner.¹⁰¹ New ways of communication through online platforms require huge efforts to protect the freedom of expression and information, as enshrined in Article 11 of the EU Charter for Fundamental Rights,¹⁰² and the need to fight hate speech that is not protected by free speech and may interfere with the enjoyment of other fundamental rights.

95 Ibidem, page 17

96 Ibidem, page 14

97 EU Fundamental Rights Agency, *Getting the future right – Artificial intelligence and fundamental rights*, 2020 [LINK](#)

98 Ibidem, page 20

99 Ibidem

100 See for example the project [Assessing high-risk artificial intelligence](#) | European Union Agency for Fundamental Rights (europa.eu)

101 See [Online content moderation – harassment, hate speech and \(incitement to\) violence against specific groups](#) | European Union Agency for Fundamental Rights (europa.eu)

102 [Article 11 – Freedom of expression and information](#) | European Union Agency for Fundamental Rights (europa.eu)

Policy recommendations

When it comes to the implementation of policy interventions to prevent the spread of borderline content, it is important to take into account and to avoid possible implications that such measures could have on **fundamental rights**.

The EU Fundamental Rights Agency (FRA) has delivered and developed projects focussing on the dissemination of online hatred against specific targets, such as minorities, religious groups, and the LGBTIQ community. As regards borderline content in relation to radicalisation and violent extremism, FRA underlines the importance of considering limitations that may be imposed on mitigating actions, as they may not exceed those provided for in **Article 10(2) of the European Charter of Human Rights (ECHR)**, without prejudice to any restrictions which Community competition law may impose on Member States' right to introduce the licensing arrangements referred to in Article 10(1) of the ECHR¹⁰³

In relation to hate speech online¹⁰⁴, which has been addressed extensively in this handbook, it is crucial to note that 'freedom of expression constitutes one of the essential foundations of [a democratic] society, one of the basic conditions for its progress and for the development of every man. Subject to paragraph 2 of Article 10 [of the European Convention on Human Rights], it is applicable not only to 'information' or 'ideas' that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broad mindedness without which there is no 'democratic society'. This means, amongst other things, that every 'formality', 'condition', 'restriction' or 'penalty' imposed in this sphere must be proportionate to the legitimate aim pursued.'¹⁰⁵ In addition, **'...tolerance and respect for the equal dignity of all human beings constitute the foundations of a democratic, pluralistic society. That being so, as a matter of principle it may be considered necessary in certain democratic societies to sanction or even prevent all forms of expression which spread, incite, promote or justify hatred based on intolerance...**, provided that any 'formalities', 'conditions', 'restrictions' or 'penalties' imposed are proportionate to the legitimate aim pursued.'¹⁰⁶ As shown in the case studies presented in this handbook, subtle forms of hate speech that are not openly inciting to violence nor bear clear links to violent extremism are being disseminated online.

103 See Article 11 – Freedom of expression and information | European Union Agency for Fundamental Rights [LINK](#)

104 The case law of the European Court of Human Rights (ECtHR) applies with an overview provided in the ECtHR fact sheet on Hate speech [LINK](#)

105 ECtHR, *Handyside v. the United Kingdom* judgment of 7 December 1976, § 49

106 ECtHR, *Erbakan v. Turkey* judgment of 6 July 2006, § 56

An important part of this handbook is dedicated to existing regulations and companies' internal policies that may have mitigating effects in the spread of borderline content leading to violent extremism and terrorism. In this context, the Institute for Strategic Dialogue in the 2022 EUIF workshop on algorithmic amplification and borderline content, has underlined how responses and mitigation measures must take into account the fact **that the borderline content is shifting** – with the threat less explicitly 'group' based and increasingly intersecting with other information harms like **conspiracy theories**. Extremists are also adopting an ever more **transnational** perspective. These factors, according to ISD, requires a new generation of responses, which go beyond existing counter-extremism policy paradigms. This includes:

- **Regulation:** Consider human-rights based approaches to social media regulation which focus on systemic risk mitigation and transparency rather than content removal alone, recognising platform design issues which facilitate the growth of extremist movements.
- **Intervention:** Develop the next generation of offline and online interventions to address these looser movements and bringing people out of extremist ecosystems.
- **Prevention:** Formulate programmes, including curricula and community initiatives, required to raise awareness of and build resilience against hybrid extremism threats. Greater efforts should be made to educate publics and develop critical thinking, including partnering with the private sector, schools and universities, religious and youth organisations.
- **Coordination:** Facilitate improved regional policy exchange, recognising the transnational nature of these threats, and consider lessons learned from mechanisms developed for international collaboration to counter specific groups (e.g. Counter Daesh Coalition).

CCDH Policy Recommendations on preventing the spread of violent misogynist content

In the context of preventing the spread of **misogynist content** leading to gender-based violent extremism, the Centre for Countering Digital Hate has developed a framework of action called 'the STAR Framework' for legislative design. The idea behind the framework is that all recommended interventions will be enhanced and embedded if supported by a robust regulatory framework. CCDH has designed the STAR framework to support global efforts to regulate social media and search engine companies, and ensure consistency, effectiveness and connectedness for global problems, like Incels and other forms of violent extremism.

The STAR framework has four key components:

- Safety by Design
- Transparency of Algorithms, Rules Enforcement and Economics
- Accountability to Independent and Democratic Bodies
- Responsibility of Technology Companies and their Senior Executives

And it includes:

- Transparency and enforcement of a platform or search engine's rules with easy complaint pathways and responsiveness; and
- Independent / democratic accountability structures with real world consequences for both the companies and the senior executives in order to sustainably change corporate behaviour

Conclusion

The information, analysis and guidance gathered in this handbook confirm that the establishment of **clear standard definitions and measures** to limit the spread of borderline content in relation to violent extremist is challenging due to the different business models and formats of online platforms, as well as the need for preventing any breach of fundamental rights, such as freedom of speech.

Nevertheless this handbook intends to provide the basis to reach a common ground of understanding of the type of **content and tactics** that violent extremists and terrorists are using to evade detection and to increase the reach of their propaganda campaigns and narratives.

Borderline content and tactics in the context of the handbook can be summarised as such:

- 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).
- **Tactics** used to **manipulate users** and **amplify** borderline content leading to violent extremism, such as algorithmic amplification techniques that profit from biases in content sharing algorithms.

The involuntary spread of borderline content leading to violent extremism by online platforms' recommender systems has also been identified as important element for the scope of this handbook.

Information on borderline content provided by Member States and Civil Society Organisations has highlighted the predominance of the following **categories** of either illegal content, which can be challenging to detect as such, or legal but harmful content:

- **Content** that targets migrants and specific religious groups, such as the Jewish community and Muslim people.
- **Anti-LGBTIQ** and misogynist content.
- **Conspiracy theories and misinformation** concerning the war of aggression against Ukraine and the COVID-19 pandemic (including vaccination campaigns).
- **Anti-government/system** content meant to incite violence.

Recommendations to online companies suggest the need for:

- More and better **transparency** and **risk assessment** processes, which will be facilitated with the implementation of the DSA in the EU, and access to data to trusted researchers.
- Better means to identify **hate speech** and link it to violent extremism.
- Better content moderation tactics that go beyond removal and that focus on **deranking** and **demonetising** legal but harmful content that can lead towards violent extremist and terrorist acts.
- Better measures to avoid **algorithm biases** that can lead to the automated spread of hate speech and legal but harmful content related to violent extremism and radicalisation.

Stakeholders agree on the need to address **gaps and challenges** encountered by tech companies, especially small ones, and governments in establishing **common definitions** and in assessing when **thresholds** in the breach of terms of services and community guidelines are met.

The EUIF, GIFCT, and other stakeholders involved in this work, have provided a series of examples and measures that have been taken so far by both the tech industry and EU Member States to support the development of future practices that should be discussed and implemented in the EUIF and other relevant multi-stakeholder networks and fora. The EUIF intends to keep this handbook up-to-date, as a means to continue supporting tech companies in addressing challenges related to content moderation.

A conclusive note by GIFCT

As 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 multistakeholder 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, processes of radicalisation or activities 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. As 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. The more sub-themes relate to real world harm, the more likely clear remedial actions can and should be taken by tech companies. Overarchingly, the more broadly a sub-theme aligned with 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.

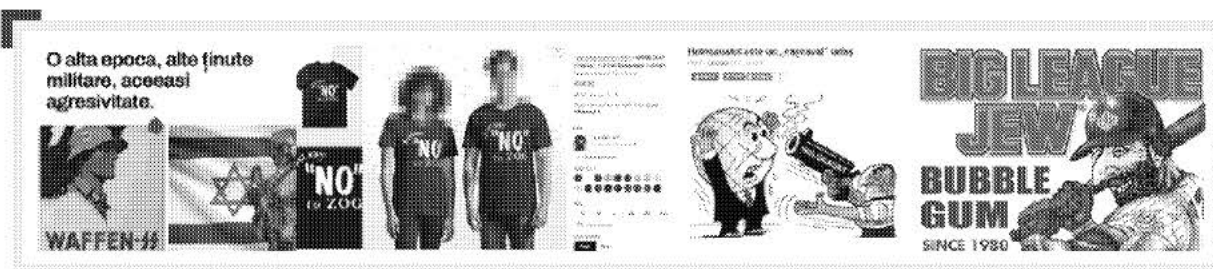
Government Guidance by GIFCT: The more governments can define the TVEC related harm areas they are most concerned about, and the more this can tie to legal frameworks, the easier it is to encourage actions by tech companies in a principled manner. Even in cases where content is not removed but is downranked or demonetised by tech companies, there need to be principled policies behind the actions that are definable, defensible, 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 Multistakeholder 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 multistakeholder 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.

Annex I

Handbook's Glossary

Antisemitic slurs and terms



Antisemitic imagery: Holocaust denial and distortion – Is a discourse and propaganda that attempt to deny the historical reality and the extent of the extermination of the Jews by the Nazis – a belief that the Holocaust did not happen or was greatly exaggerated. Moreover, it is an intentional effort to excuse or minimise the impact of the Holocaust or its principal elements, minimising the number of the victims of the Holocaust in contradiction to reliable sources and blaming the Jews for causing their own genocide.

International Holocaust Remembrance Alliance (IHRA) non-legally binding working definition of antisemitism (including its examples): 'Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities.'¹⁰⁷

ZOG: is the acronym for 'Zionist Occupied Government'. It refers to a far-right conspiracy theory reflecting the idea that the government is controlled by Jews.

JWO: is the acronym for 'Jewish World Order', the antisemitic version of the New World Order conspiracy theory, claiming that the single world government will be lead by Jewish people.

NOW: is the acronym for 'New World Order' a conspiracy theory that argues that a shadow elite is trying to implement a totalitarian world government.

Some of this propaganda is supported by the dissemination of imagery as the example below shows:

Rothschild(s): The Rothschild family is a wealthy Ashkenazi Jewish family originally from Frankfurt that rose to prominence with Mayer Amschel Rothschild (1744–1812), who established his banking business in the 1760s. The Rothschild family has frequently been the subject of conspiracy theories, many of which have antisemitic origins

Jewish Agenda: It is a conspiracy theory that claims that a malevolent, usually global Jewish circle, referred to as International Jewry, conspires for world domination.

¹⁰⁷ [What is antisemitism? | IHRA \(holocaustremembrance.com\)](https://www.holocaustremembrance.com/what-is-antisemitism)

Anti-migrants and Islamophobic slurs and terms



Countryballs/Polandball: These are very widespread memes mainly used as form of political caricature. It consists in drawings on balls representing different countries with stylised faces. This kind of meme is often used to comment on current political events and does not necessarily have violent right-wing extremist references.

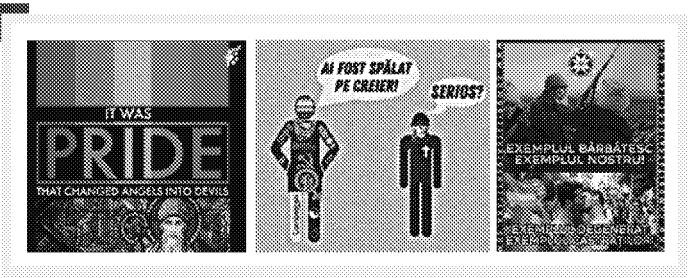
Superspreaders: Online actors responsible for the widespread distribution of harmful content.

Online subculture of the Alt-Right: The alt-right is an abbreviation of alternative right, which is a far-right, white nationalist movement with a largely online community.

Anti-migrants and Islamophobic content: The anti-migrant’s discourse is the opposition to immigrants or immigration in general, being characterised by or expressing opposition to or hostility toward immigrants. As for the islamophobia, it is the fear of, hatred of, or prejudice against the religion of Islam or Muslims in general, especially when seen as a geopolitical force or a source of terrorism. The use of disinformation, manipulation techniques and sharing memes have also been popular among anti-migrant extremists, as the follows:

Dehumanising Speech: Speech or language that labels a group based on a protected attribute as biologically subhuman (‘cockroaches,’ ‘microbes,’ ‘parasites,’ ‘yellow ants’), mechanically inhuman (‘logs,’ ‘packages,’ ‘enemy morale’), or supernaturally alien (‘devils,’ ‘Satan,’ ‘demons’).

Anti-LGBTIQ and incels-related slurs and terms



LDAR, or ‘Lay Down And Rot’: self-harm and suicidal frequently used on incel forums.

Roastie: It’s a slang term used by the incel’s community to target and disparage sexually active women.

Groomer: It’s a slang term used in the anti-LGBTIQ communities to refer to an adult who wants to exploit or abuse impressionable children, either sexually or otherwise, typically by building trust and an emotional connection with them.

Roastie: Slang used by the incel community to target and disparage sexually active women

Dyke/Lesbo: A term originated as a slur against more masculine-presenting lesbian women, this term has been reappropriated by its community into being a common slang term to refer to lesbian women. While some people may not mind being referred to with this term, stakeholders should be aware that the term can also carry negative connotations and many people may be uncomfortable with the term.

Thing: Specifically in reference to pronouns, the use of 'thing' instead of a user's preferred pronouns. It is used to usually mock a user's preferred way of expressing their gender identity. It is commonly used as a way to invalidate or minimise trans/enby people.

Fag/Faggot/Homo: All terms used to refer to gay people, and all heavy slurs with the intention of belittling and attacking this group. These words are also commonly used in the real world homophobic attacks, both verbal and physical. These terms should not be taken lightly.

Trap: A term that originated from anime. This word is in reference to men who dress as women and are feminine-presenting, and 'trap' heterosexual people into having an attraction to them. This word has been used out of its original contexts as a slur to transgender people, as if their existence is to 'trap' or 'trick' people around them. Not everyone finds this term offensive. It should be evaluated as to whether it deserves moderation on a case-by-case basis.

Sigma: a pseudo-scientific construct of the hyper masculine alt-right denoting a particular type of male behaviour (successful, highly independent, intelligent). In effect, 'sigmas' are an equivalent of 'lone wolves'. As reported by the Slovakian Council, currently, the hyper masculine alt-right, best represented by the likes of Andrew Tate, is obsessed with the personality of Patrick Bateman from the movie American Psycho which, according to the community, best describes the qualities of 'sigmas'.

Anti-feminist: It's the opposition to some or all forms of feminism, opposing to particular policy proposals for women's rights and has sometimes been an element of violent far-right extremist acts.

Misogyny: Misogyny is hatred of, contempt for, or prejudice against women. It is a form of sexism that is used to keep women at a lower social status than men, thus maintaining the social roles of patriarchy.

Anti-LGBTIQ: It is the movement that opposes to LGBTIQ legal rights, particularly to civil unions or partnerships, LGBTIQ parenting and adoption, military service, access to assisted reproductive technology, and access to sex reassignment surgery and hormone replacement therapy for transgender individuals.

Incels: Are members of a group of people (men) on the internet who are unable to find sexual partners despite wanting them, and who express hate towards people whom they blame for this, especially women.

Red pilling: Term used to describe the so-called realisation that men do not have power or privilege in society. Contrary to this, men are vulnerable to being exploited by women in social, economic and sexual terms.

Black pill: Theory alleging that the ability of an individual to establish romantic relationships is determined by appearance and therefore genetics. It is also used in the context of thinking that all is doomed; that ZOG has already won and that their race will never see victory.

White pill: Word used to stress their confidence in their race victory over others.

Chad: Term used to designate men who are presumed to be sexually desirable and therefore popular amongst women.

Racist slurs

Jap: Used during World War II when the Japanese were in internment camps in the US, this term was a derogatory way US citizens referred to Japanese people and is heavily considered an ethnic slur against Japanese people.

Gypsy/Gypped: Both to refer to 'Gypsy' the people, and to be robbed/conned in the form of 'gypped', this is a term specifically used as an ethnic slur against the Romani people. While it is used in legal contexts, the words have slowly been brought out of use due to its common use as a slur historically.

Chink/Ching Chong: Chink has been historically used as a slur against people of Chinese descent, and sometimes even Asian decent widely, with ching chong mocking the language of the Chinese which is commonly used alongside chink.

Triple Parentheses, also known as (((echo))): This is a very uncommon but recently used symbol to denote someone of Jewish origin, typically in a way to target or harass them. This symbol is used to single Jewish people out by communities and places a target on their back for their religion or ethnicity, and should not be tolerated.

Curry: Often used in an offensive way against Indian community.

Noodlewhore: The word used by incels to refer to an Asian woman.

Kike: It is an ethnic slur for a Jew.

Noodle: Term used to insult Southeast Asian communities.

Currycel: Name used to refer to an Southeast Asian incel.

Monkey: Term used by incels to mock and discriminate African people.

Sheboon: It is a racist term for an African American woman. Combination of 'She' and 'Baboon'.

Ape: It is an offensive term used by extremists to describe a non-white person.

Sand Nigger: Ethnic slur used against black people, especially African Americans.

Spic: It is an ethnic slur term used to referring to people from Latin America.

Paki: Term used to refer to a person from Pakistan. Can be used as an insult.

Negroid: Slur term referring to an African person.

Negro: In certain contexts, it can be used as an insult against African communities.

Abo: Offensive slang used as a disparaging term for an Australian Aborigine.

Mongoloid: A derogatory term for a mentally challenged person. This used to be used to denote people with Down's Syndrome, and now is just generally derogatory.

Faggot: Often shortened to fag, is a usually pejorative term used to refer to gay men.

Fag hag: It is used against women that they claim like to be to associated with androphilic men.

Tranny: It is an offensive and derogatory slur for a transgender individual.

Trannie: It is an offensive and derogatory slur for a transgender individual.

Queen: It is the slur used against a homosexual man who does not have the stereotyped characteristics of a man.

Dyke: Slur term used against lesbian women. It is often used by homophobic women.

Fairy: Term used in an offensive way to men that incels consider less masculine than them.

Lesbo: Pejorative term to refer to a homosexual woman.

Sodomite: An offensive term used to describe a homosexual man.

Nancy: Term to discriminate a homosexual man, considering him an effeminate man.

Poof: A derogatory term used to describe a homosexual man.

Pansy: Term can be used as a homophobic insult which suggests a lack of manliness.

Poofter: It is one of most pejorative words in Australian English. The phrase 'poofter-bashing' arose during the 1960s and 1970s during organised hate crimes against homosexuals in Australia. It is an offensive term which considers gay men as feminine.

Fudgepacker: Slang and derogatory term to label a male homosexual.

Genderbender: It is used in a vernacular by incels way to refer to a person who dresses up and presents themselves in a way that defies societal expectations of their gender, especially as the opposite sex.

Foid: Incels use this term when calling another person an ass, asshole, or very stupid.

Landwhale: A landwhale is a pejorative term for an overweight woman that is commonly employed in various online forums (especially those connected to the manosphere – online space where incels are active).

Femoid: It is a derogatory term used in the incel community to refer to a woman. 'Femoid' comes from the contraction of the word 'female' and 'android' (robot), to emphasize the allegedly icy nature of women.

Ho: Used by the incel community to discriminate women.

Thot: Thot is a denigrative slang term used against women, originally defined as an acronym of 'That hoe over there' but now generally used as a synonym for hoe or slut, especially in the manosphere.

Hag: The term is commonly used by incels to describe what they consider as a hysterical or ugly women in positions of power.

Goblin: A racist insult that claims that other ethnicities are an undesirable and unnatural race that grows in damp and dirty places.

Gold digger: A person whose romantic pursuit of, relationship with, or marriage to a wealthy person is primarily or solely motivated by a desire for money. This term is often used in some forums to ridicule women in general.

Cum dumpster: Often used to label a sexually active women as a perverted or promiscuous person.

Skank: It is often used by incels against women and when used in this context, can be similar to prostitute.

Slag: In slang, slag is an insulting term for a contemptible person. When used against women, it can be equivalent to slut.

Annex II

GIFCT Review and Analysis of Tech Company Efforts to Counter Borderline Content

Bringing borderline content to the fore of multistakeholder 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 reviews the types of content that fall under scope of 'borderline content', and what feasible actions on content looks like in real terms for tech companies. GIFCT reviewed policies and actions taken by its 22 member companies for this analysis.

GIFCT: Variations in Tech Approaches to Borderline Content

Above 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:

- The human resources with subject matter expertise that a platform is able to hire.
- The engineering and tooling support a platform is able to give to a harm type.
- The awareness or prevalence of a certain harm type on the platform.
- 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 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 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 nor 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.

Reviewing GIFCT Member Companies Policies and Actions

GIFCT reviewed its 22 GIFCT member companies' policies against the 14 sub-themes identified as making up borderline content in relation to TVEC. As mentioned in the body of this handbook, the borderline content sub-theme policies identified by GIFCT included the following, notably mapping how each theme could manifest in relation to terrorism and violent extremism but also in non-TVEC related ways online:

1. Violent content, graphic content, gore
2. Weaponry and instructional material
3. Symbols, slogans and visual indicators associated with violent extremist groups
4. Meme subculture
5. Incitement to violence
6. Hate speech
7. Bullying, harassment, and threats
8. Anti-Refugee or anti-immigrant sentiment
9. Stereotypes and dehumanisation
10. Mis- and disinformation
11. Populist rhetoric and wider nationalism
12. Anti-government and/or anti-EU sentiments
13. Anti-Elitism
14. Political satire

It is one thing to have policies around a type of content, and another to develop tools for potential enforcement actions. The range of actions taken by tech companies listed here were drawn from the TSPA enforcement methods.

Enforcement Actions of Tech Companies in Moderation Efforts

Enforcement Action	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. Allows 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.' ¹⁰⁸

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

Importantly, 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.

108 Table provided by GIFCT as contribution to the EU Internet Forum Handbook on Borderline Content. Any feedback, questions, or comments can be emailed to enr@gifct.org and micalle@gifct.org

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 act 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.

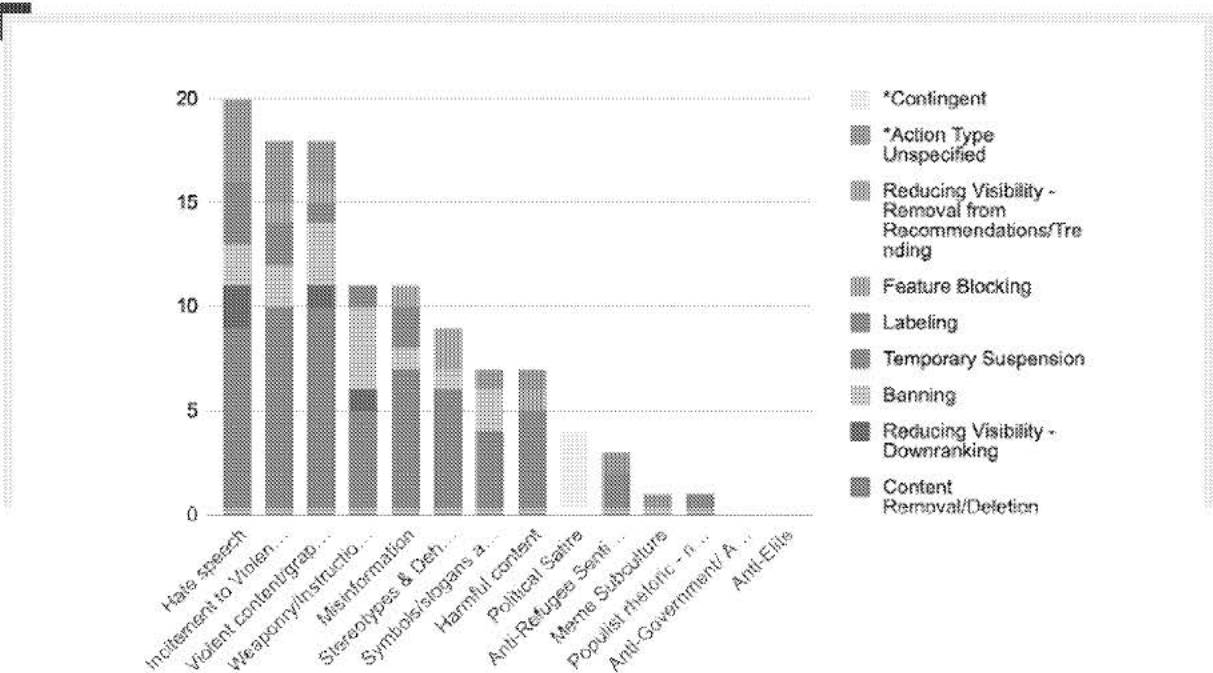
Smaller and less resourced **platforms** will be limited in their ability to undertake enforcement actions that require more **nuanced moderation or technical expertise**. Further, these smaller platforms may be unable to proactively contribute to the policy debates on what types of content should be moderated and may be more likely to follow the lead of their larger, more established counterparts.

Reviewing Company Policies and Actions

GIFCT reviewed its 22 GIFCT member companies’ policies against the 14 sub-themes identified and in relation to potential moderation 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

109



The table established by GIFCT shows where and how GIFCT member companies currently take action against TVE related borderline content sub-themes. References to borderline content types and enforcement actions were gathered from publicly available Terms of Service, User Guidelines, and public statements made by companies.

109 Table provided by GIFCT as contribution to the EU Internet Forum Handbook on Borderline Content. Any feedback, questions, or comments can be emailed to en@gifct.org and info@gifct.org

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 Action on Hate Speech,¹¹⁰ while the European Union has undertaken a number of efforts to act against hate speech¹¹¹ 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 Article 20, paragraph 2 of the International Covenant on Civil and Political Rights (ICCPR).¹¹² Accordingly, 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. Additionally 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.

¹¹⁰ United Nations, *United Nations Strategy and Plan of Action on Hate Speech*, May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

¹¹¹ 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/combating-discrimination/racism-and-xenophobia/combating-hate-speech-and-hate-crime_en

¹¹² United Nations Human Rights Office of the High Commissioner, *International Covenant on Civil and Political Rights*, United Nations OHCHR Official Site, 16 December 1966, <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

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 widely protected by democratic principles on speech.

All GIFCT members protected broad user rights to be critical of political structures, governments or elites, in line with understandings of protected free speech in democratic environments. Types of content that are heavily context dependent or ill-defined, including meme subculture or the usage of symbols, slogans, and visual indicators associated with violent extremist or non-violent extremist groups were also less likely to be actioned without some indication of context. Memes in particular are incredibly context-specific and rapidly evolve, which makes these visual indicators hard to regulate.. GIFCT members reviewed all had high levels of protection for speech where clear connections to harm was not apparent.

As discussed within the wider handbook, it will continue to be important for governments, tech companies, and experts to come to agreements and understanding about the framing of borderline content. This includes non-tech company stakeholders ensuring they better understand the policies and practices already in place by a variety of tech platforms as well as for tech companies to better understand how international government guidance and regulation is evolving to have an effect in this space. Multistakeholder partnerships and forums will remain critical to ensure this understanding and continued knowledge sharing.

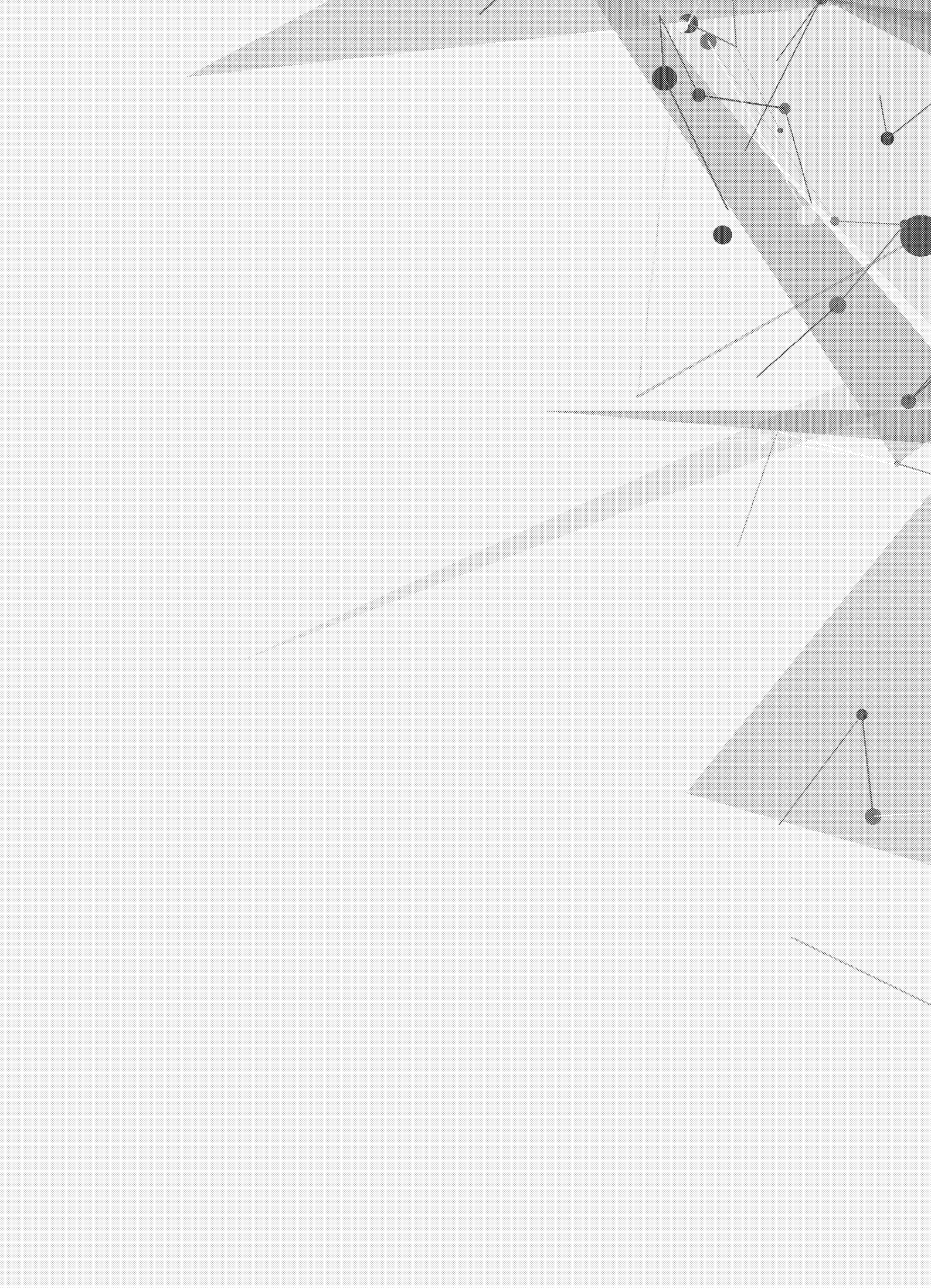


Exhibit 39

Study on the Role and Effects of the Use of **Algorithmic Amplification** to spread **Terrorist, Violent Extremist and Borderline Content** across leading **Social Media sites in Europe.**

Sponsored by the European Commission, DG Home Unit

Tom Siegel | **CEO & Co-founder - TrustLab**

Peter Dudič | **Director - Public Sector Europe - TrustLab**

Theos Evgeniou | **CIO - Tremau, Professor - INSEAD**



Berlin | London | San Francisco | Singapore





The Study.

Study Goal.

To Inform the EU Internet Forum with Measurements on the **role and effects of the use of Algorithmic Amplification** to spread Terrorist and Violent Extremist and Borderline Content **on social media platforms, across EU Member States.**

28.07.2022 - 04.05.2023



Project Sponsor.



**The European Commission DG Home Unit Directorate
Migration and Home Affairs.**

**Security Directorate at EU Prevention of Radicalization
Unit.**

Yolanda Gallego-Casilda Grau

Head of Unit Prevention of radicalisation



Tom Siegel

Peter Dudič

Xiaolin Zhuo

Nick Miller

Fabienne Meijer



Anna Maria Carpani



Prof. Theos Evgeniou

Co-Founder and CIO



Scope of Study.

5
Platforms



FACEBOOK



INSTAGRAM



TIKTOK



TWITTER



YOUTUBE

40
Weeks



8
Languages



ARABIC



ENGLISH



FRENCH



SPANISH



GERMAN



ITALIAN



POLISH



RUSSIAN

Key findings.

1. Amplification Confirmed

TVE Amplification was confirmed on all platforms. All recommend TVE content after users engage. And the filter bubble increases with continued interaction.

4. Limited Follow-up Removal

If Platforms don't catch TVE content immediately, they are unlikely to remove it proactively at a later point in time

2. Platform Experiences Differ

Stark differences between platform behavior, with Twitter and Italian often on the higher end, and TikTok and Arabic on the lower end of exposure and amplification metrics

5. Absence Of Direct Data Access Is Burdensome

Without direct data access to platforms, data collection is a time and resource intensive process, with more limited insights

3. Platforms Respond to Scrutiny

Platforms do better when they face increased public scrutiny such as Right Wing TVE Arabic Language, etc

Motivated New Users.

Human Agents | Random Personas

Characteristics

- 👤 New to the platform
- 👤 Interest in TVE content
- 👤 Knowledgeable about TVE
- 👤 Searching only for TVE

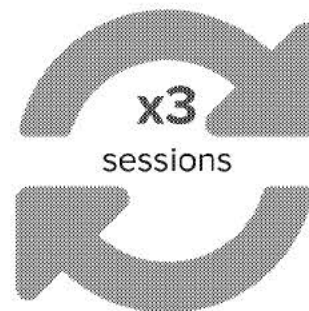
Restrictions

- 👤 Keyword list
- 👤 Interactivity: Low / High
- 👤 TVE Type: Left / Right / Int.
- 👤 Time limited search

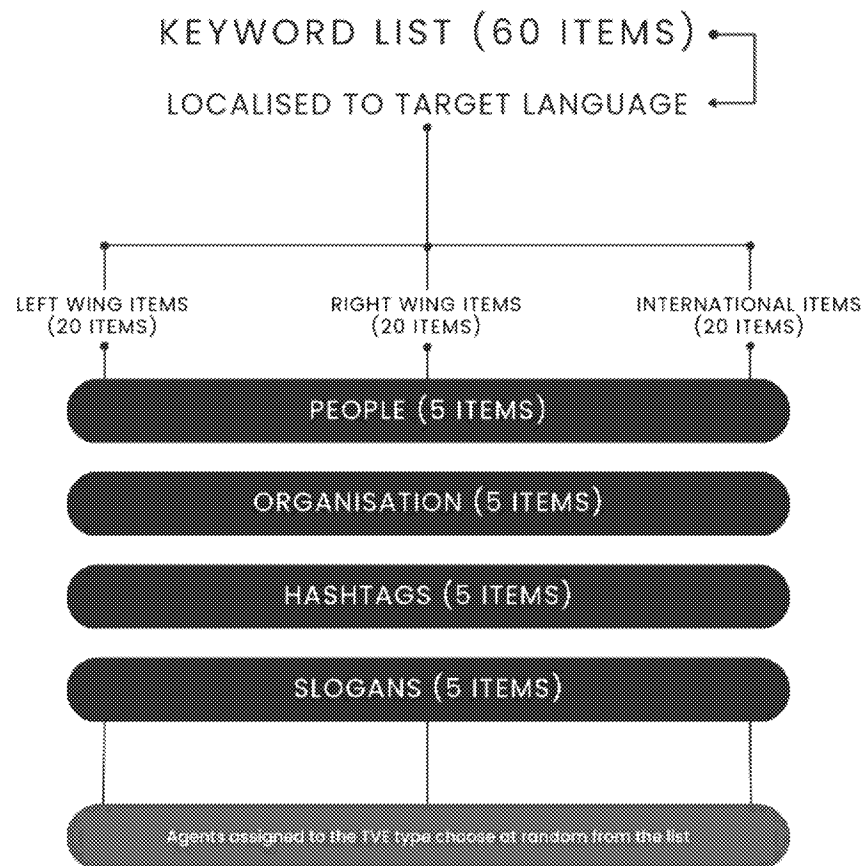
Measurement Scope.

Any media, including text, images, and videos, that **promotes or glorifies terrorism or violent extremism**, or advocates for the use of violence to achieve political, ideological, or religious goals.

Interact
Find TVE



Evaluate
Collect Feed



Datasets.

Collection



Interaction Dataset

The sessions and the TVE content found by agents



Evaluation Dataset

The content of the profile's feed after each session



Analysis



Findability Dataset

Cleaned copy of Interaction Dataset

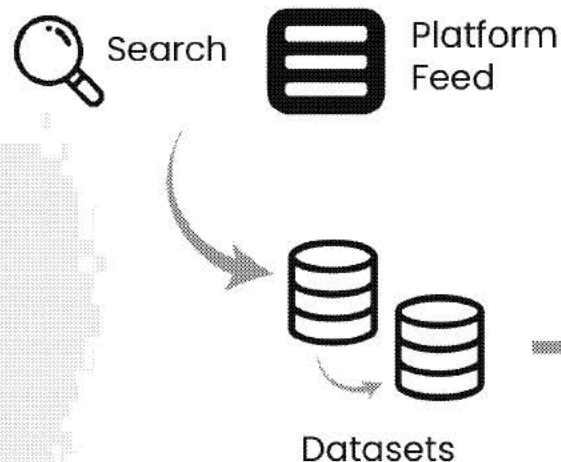


Sample Dataset

Random uniform sampling of content across both datasets, cleaned and labeled

TVE Labeling.

Categorizing content



Bad Topic

Is the content related to TVE regardless if it is actually TVE.
Includes EDSA.

Bad Content

Promotes or glorifies terrorism or violent extremism

Borderline

Not explicitly promoting or supporting or TVE that doesn't break platform policies.

Key Takeaways and Metrics.

Key Metrics:

Dissemination of TVE content.



Findability

How much TVE content can a motivated actor find in a one hour session?



Action Rate

How much TVE content is actioned by the platform?



Time to Action

How long does it take a platform to label or remove such content?

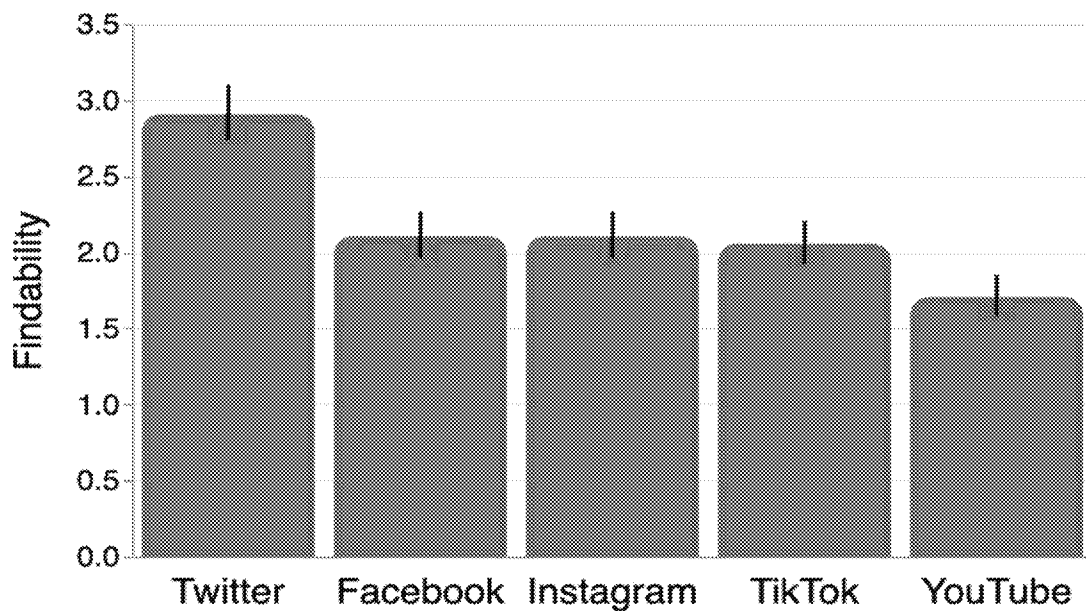


User Sentiment

What is the user perception regarding such content?

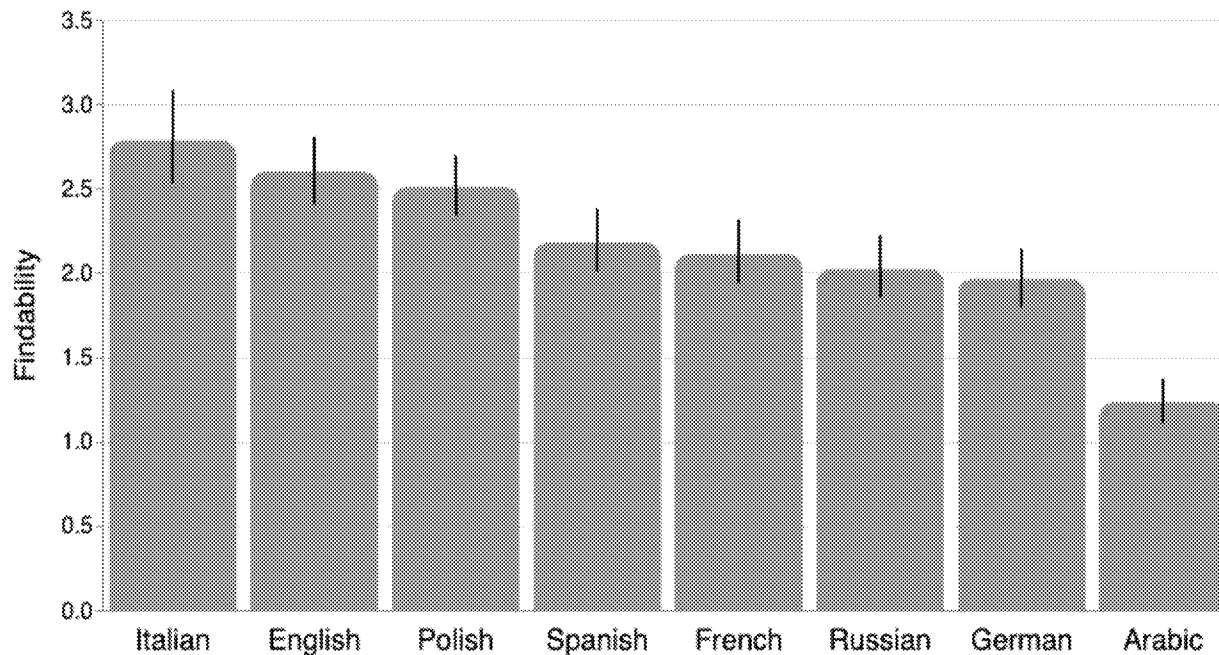
TVE Content Findability.

Most TVE Content was found on **Twitter**, and in **Italian**.



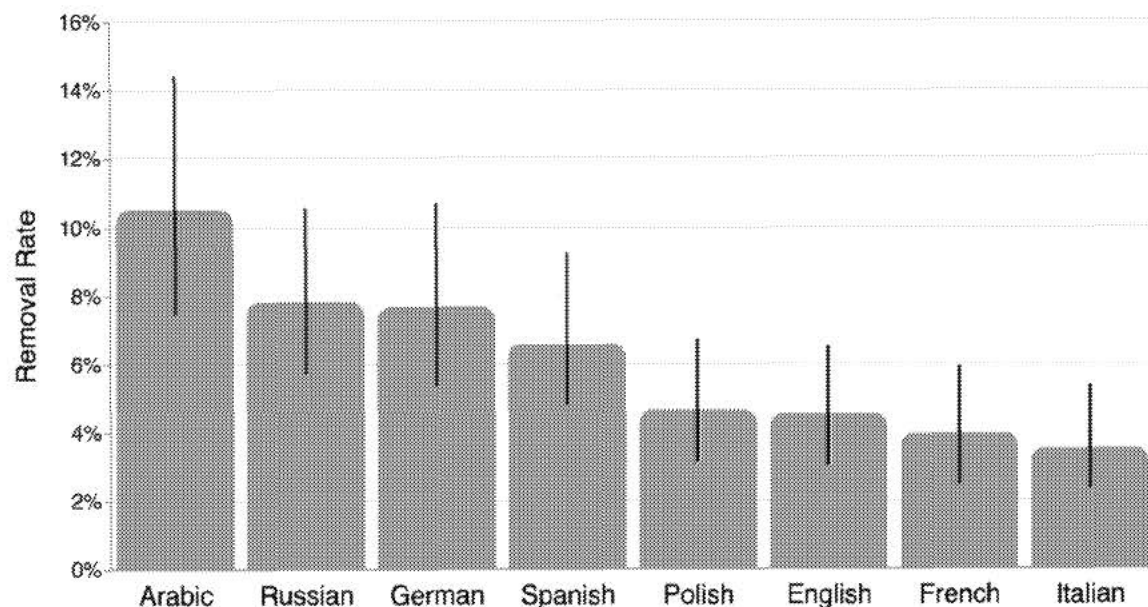
TVE Content Findability.

The least TVE Content was found in **Arabic**, and on **Youtube**.



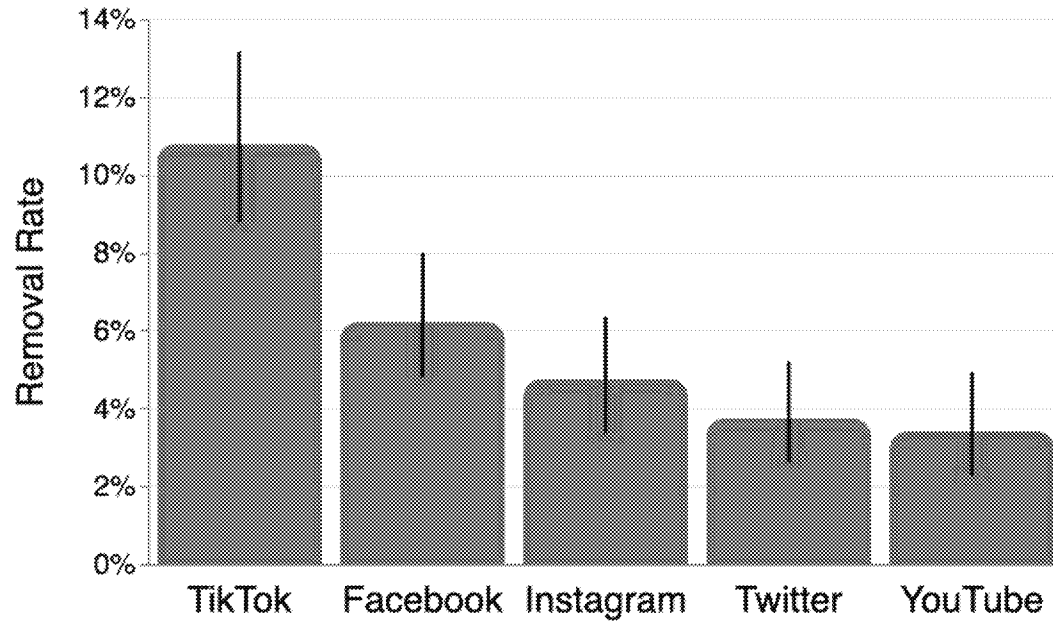
TVE Content Removal Rate.

Arabic language TVE content was proactively removed at the highest rate.



TVE Content Removal Rate.

TVE content on **TikTok** was proactively removed at the highest rate.



Key Metrics.

Automated Dissemination of TVE



Amplification

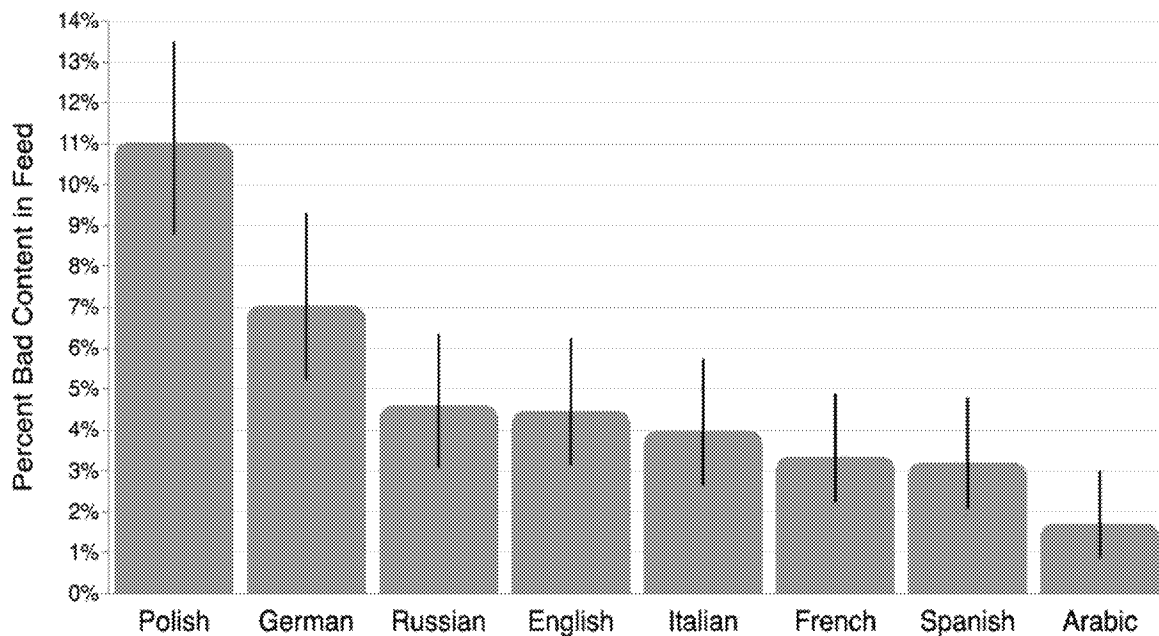
What percentage of the platform feed is associated to harmful content.

Bad Content

Borderline

Content Amplification.

Content in **Polish** language, on **Twitter** and **Left Wing TVE** get amplified the most. Amplification reaches **up to 11%**.



4.91%

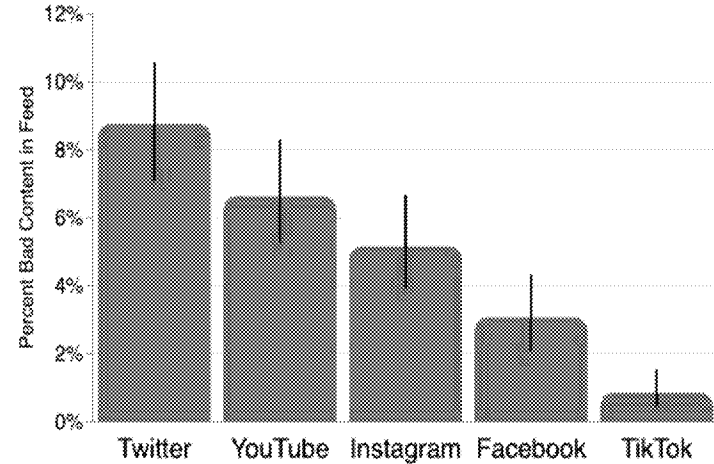
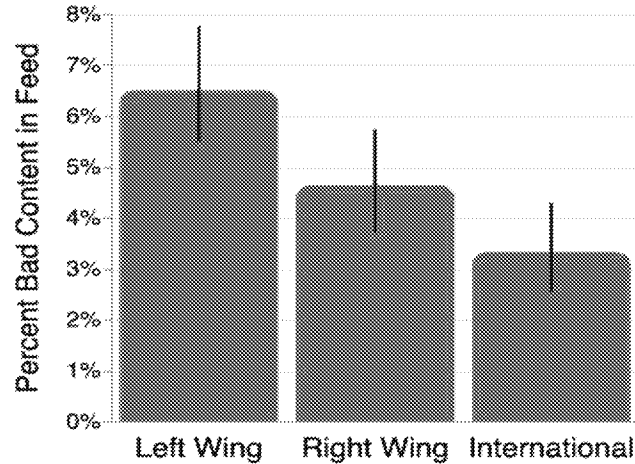
Average Amplification

Content Amplification.

Content in **Polish** language, on **Twitter** and **Left Wing** TVE get amplified the most. Amplification reaches **up to 10%**.

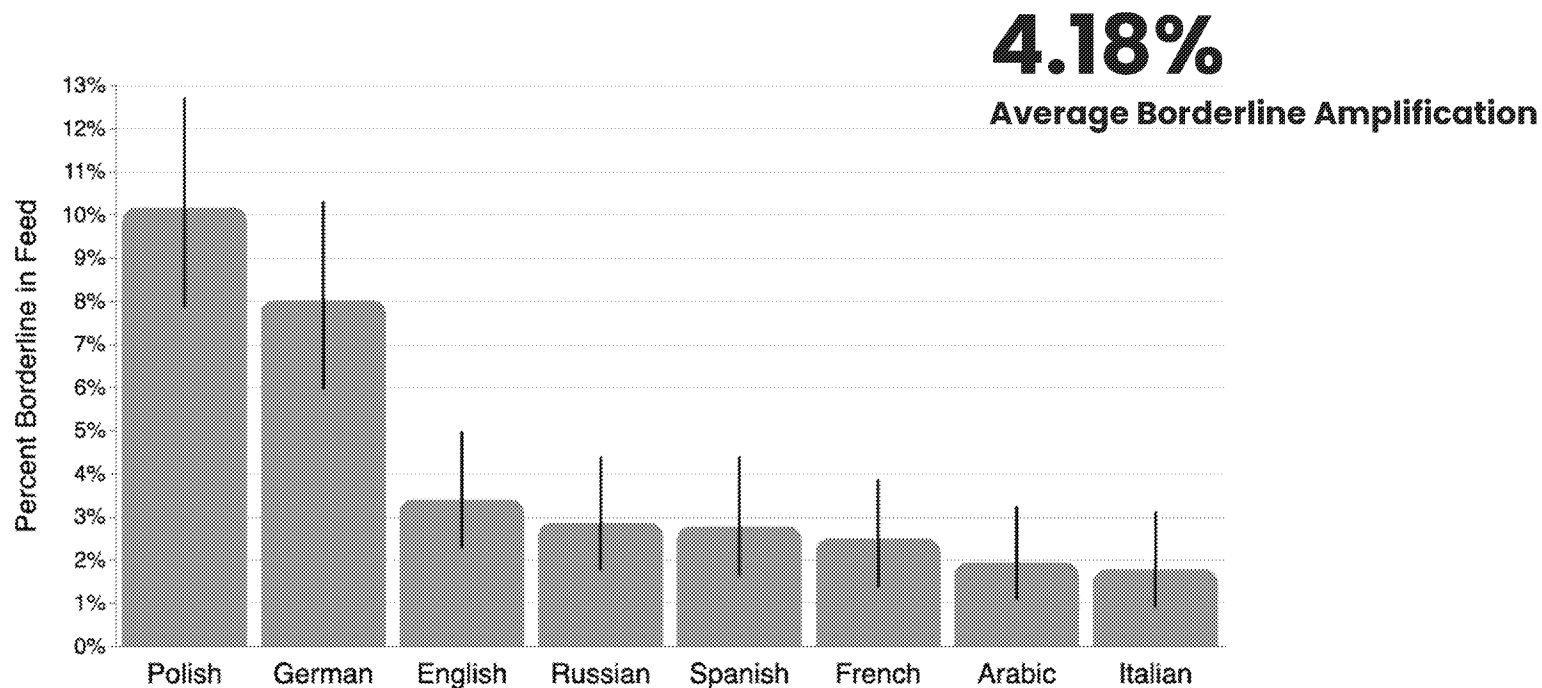
4.88%

Average Platform Amplification



Borderline Content Amplification.

Borderline content amplification patterns are is similar to TVE Content, with **Polish**, **German**, **Youtube** and **Left Wing TVE Content** leading.

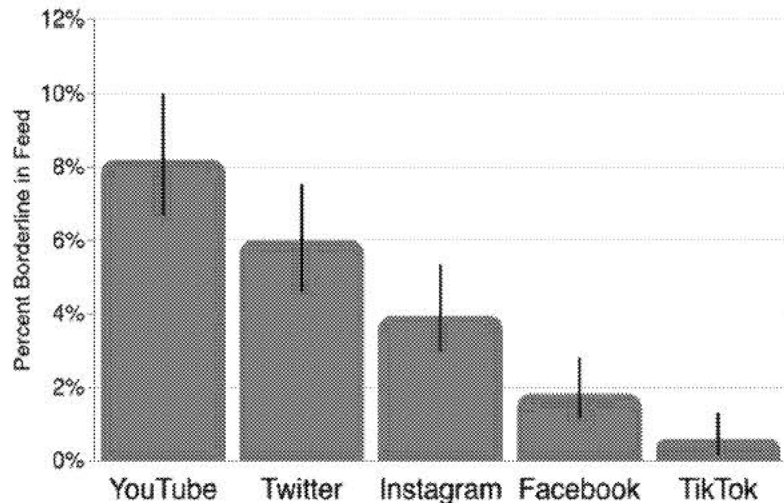
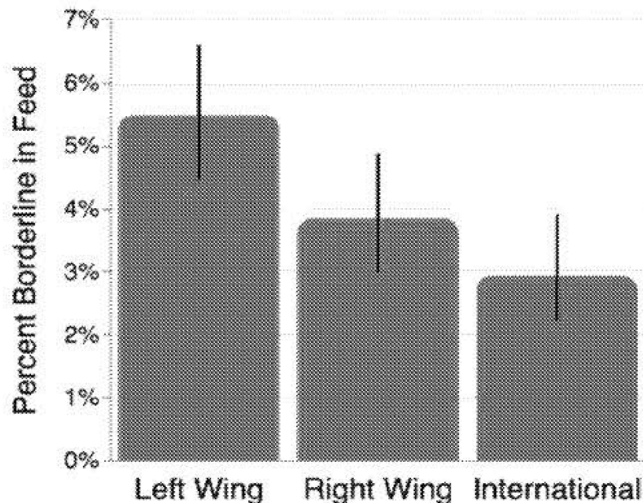


Borderline Content Amplification.

Borderline content amplification patterns are is similar to TVE Content, with **Polish**, **German**, **Youtube** and **Left Wing TVE Content** leading.

4.10%

Average Platform Amplification



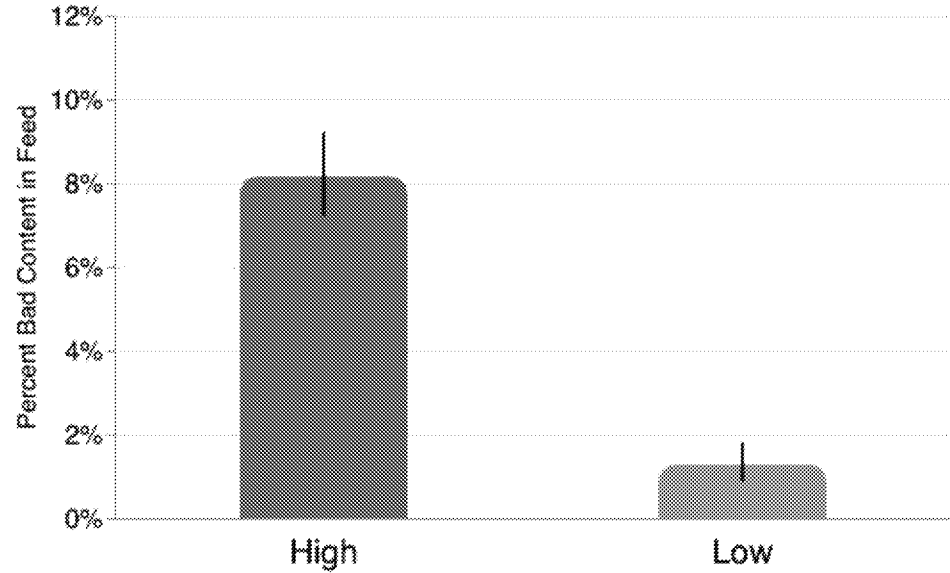


Other interesting insight.

Interactivity.

Interaction impacts amplification

Higher **interaction**
with content results in
content amplification



Amplification Across All Platforms
by Interaction Level



Recommendations.



Study Recommendations.

1. Platforms need to do more

A lot more work needs to be done to reduce the spread of TVE content. Examples and data from this study can help.

2. Need for Additional Data deep dives & analysis

Platforms, researchers and the Commission should come together to assess this dataset in more depth.

3. Importance of continued measurement

To effectively reduce TVE in social media, it needs to be continuously measured by independent 3rd parties.

4. Improved Data access for scientific research

With direct access to platform data and more time and budget, deeper and more reliable insights are possible.

5. Borderline Content needs more attention

Risk and impact need to be better understood and guardrails defined, to limit its impact and contributions.

6. Recommend Monitoring and Prevention

IT may be possible to predict and prevent or improve the impact on specific groups/content/context

Deep Dive Analysis.



Some Key Questions

- ❑ Can platforms predict what/who/why is amplified (and manage it preemptively)?
- ❑ Can we understand the dynamics of amplification?
- ❑ Can we develop tools for these?

*Beware of Causality

Dynamics of Amplification

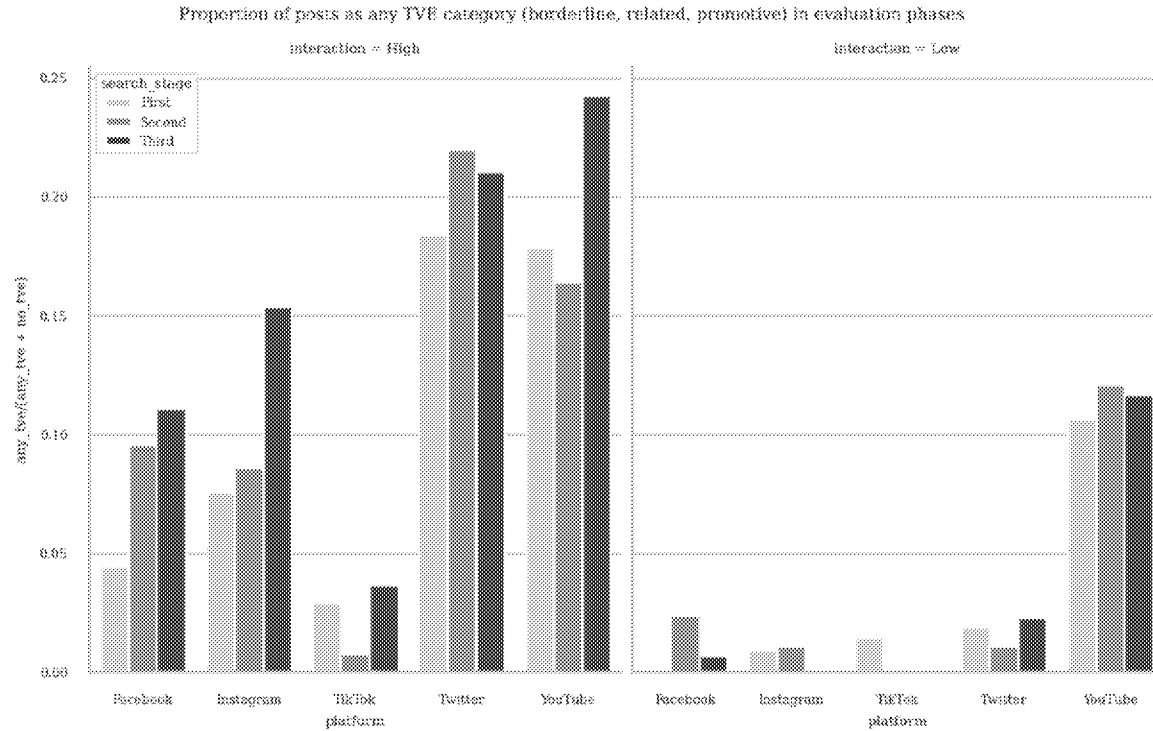
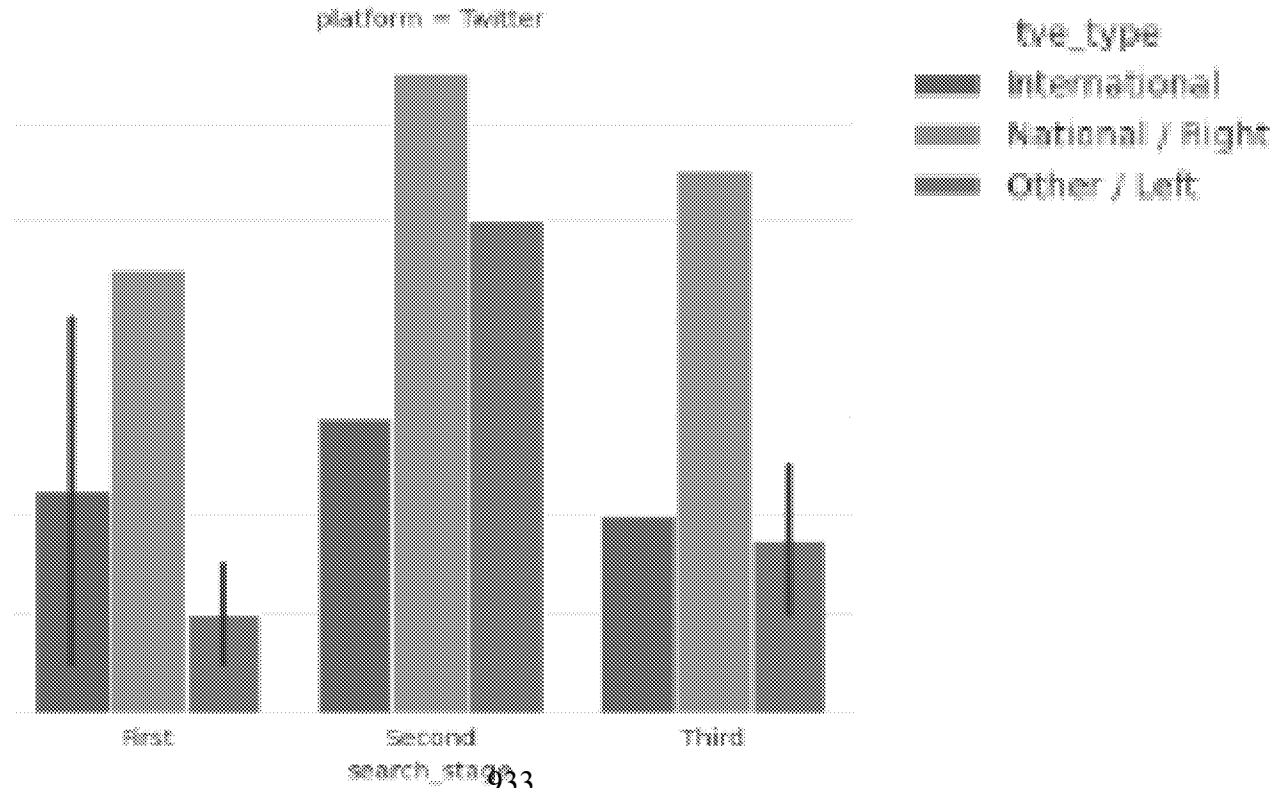


Figure 4.5.2a - Prevalence of all TVE categories, grouped by platform and divided between high and low levels of persona interaction

Dynamics of Amplification



Deep Dive Analysis: Methodology

- 1534 “Sessions”: both interaction *and* observation sessions
 - 253 have have TVE (we balanced the data via weighting)
 - Assesses each platform separately
-
- ❖ Split data in 80% training and 20% testing; 10-fold
 - ❖ Used various applicable machine learning methods:
 - Support Vector Machines; Trees, Random Forests, Gradient Boosting

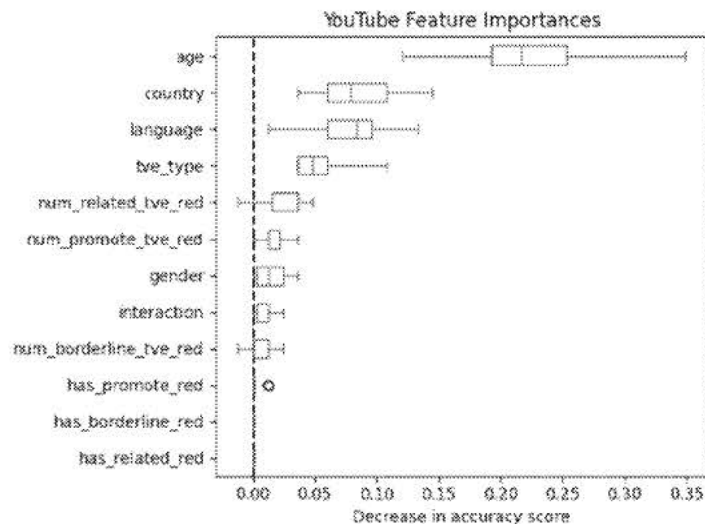
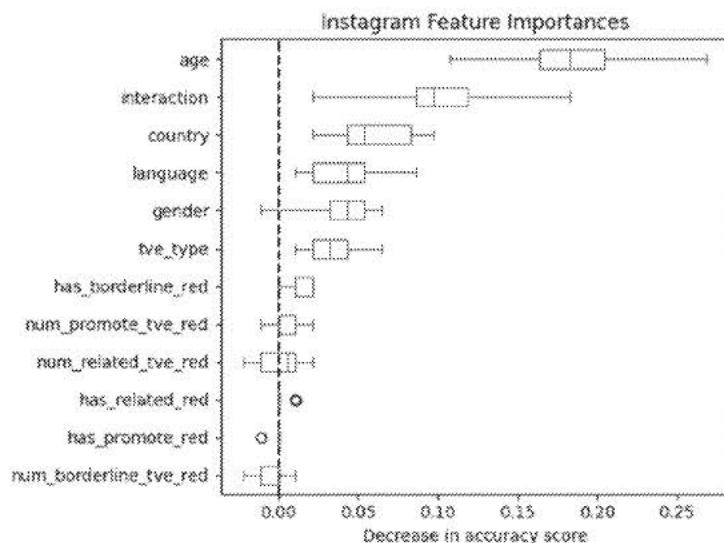
Some Features (Engineered)

- ☐ Age group
- ☐ “Risk of radicalization” (interaction X Content Type)
- ☐ Content Popularity (Likes, Comments, Shares, Followers)
- ☐ “Proximity” to TVE (relates to, promotes, borderline scores)
- + Persona’s
- ☐ Language
- ☐ Country
- ☐ Gender
- ☐ Search Stage

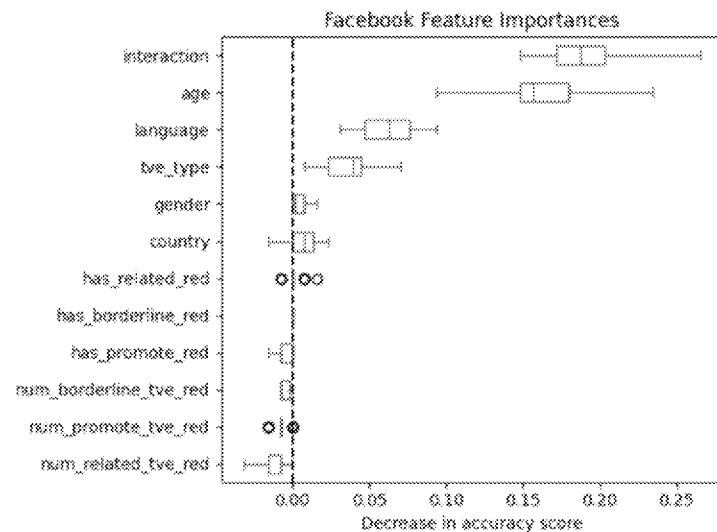
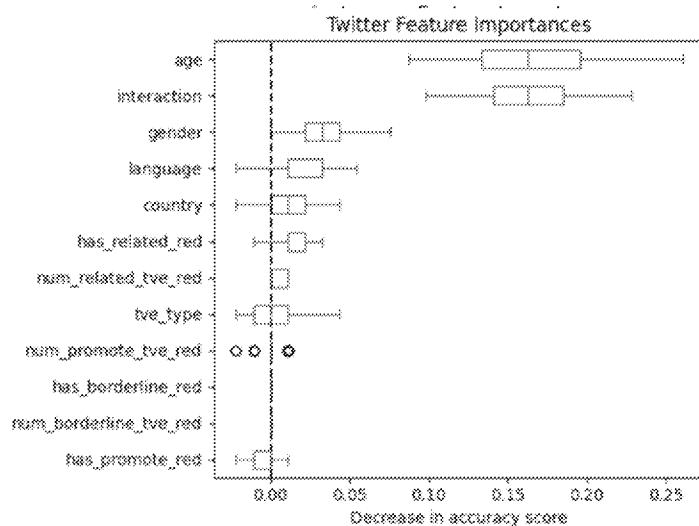
Predictability of TVE Exposure + Features to Monitor

	YouTube	Twitter	Instagram	Facebook
SV Classification	0.554	0.548	0.548	0.547
Decision Tree	0.795	0.935	0.935	0.922
Random Forest	0.807	0.925	0.925	0.922
Gradient Boosting	0.831	0.914	0.914	0.898

Predictability of TVE Exposure + Features to Monitor



Predictability of TVE Exposure + Features to Monitor



Some Key Questions

- ❑ Can platforms predict what/who/why is amplified (and manage it preemptively)? **YES**
- ❑ Can we understand the dynamics of amplification?
- ❑ Can we develop tools for these? **YES**

*Beware of Causality

Discussion.

Thank you.

Redacted

+ TrustLab

Berlin | London | San Francisco | Singapore

 TREMAU

Reports.

The report is freely available to [read and download here](#),
as well as the [Publications Office of the European Union](#).

Exhibit 40

DG HOME

EU INTERNET FORUM: STUDY ON THE ROLE AND
EFFECTS OF THE USE OF ALGORITHMIC
AMPLIFICATION TO SPREAD TERRORIST, VIOLENT
EXTREMIST AND BORDERLINE CONTENT



This is the Final Report of the study commissioned by DG HOME, drafted by Fincons, Trust Lab and Tremau, in the context of the EU Internet Forum.

Abstract

The European Commission (DG HOME) commissioned this study for the European Union Internet Forum. The study examines the degree to which the five leading social media platforms in the EU, and in particular their recommender systems which proactively present content to users, algorithmically amplify terrorist and violent extremist content to online users. The study also examines the extent to which recommender systems amplify Borderline content, such as some forms of hate speech leading to violent extremism, disinformation and other forms of legal, but harmful content, that could lead to the creation of online filter bubbles, and recruitment and radicalisation of online users. The study analyses Facebook, Instagram, TikTok, Twitter and YouTube, covering eight markets/languages in the EU. The study was conducted from July 2022 to May 2023. Prior to this Final Report, the team submitted an Inception Report and two Interim Reports that complement this report.

Executive Summary

The European Commission (DG HOME) commissioned this study, performed by Trust Lab in collaboration with Tremau and coordinated by Fincons Group, which examines the degree to which five leading social media platforms and their recommender systems¹ amplify harmful Terrorist and Violent Extremist (TVE) content and “Borderline” content² to online users.

The Report reflects the study of Facebook, Instagram, TikTok, Twitter and YouTube, that operate in eight markets/languages, namely, Arabic, English, French, German, Italian, Polish, Russian and Spanish. The study looked at: (1) the extent to which a platform’s recommender system algorithmically amplifies TVE and Borderline content to users, and the impact of the dissemination of such content on a user’s journey to radicalisation; and (2) the extent to which a platform’s recommender system leads to the creation of online “filter bubbles”³ and the impact on a user’s adoption of violent extremist beliefs.

This study was a collaboration between Trust Lab, Professor Theodoros Evgeniou, Professor of Decision Sciences and Technology Management at INSEAD, who researched the risk of radicalisation linked to the dissemination of TVE content, and Fincons Group which handled project coordination.

¹ Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content recommendations to online users. A key component of the recommender system is its recommender algorithm that determines the content a user will be served.

² As noted by the EU Internet Forum in its [Year in Review 2022](#): “Borderline content, also known as harmful, but legal content, is content that comes close to infringing on the community guidelines of social media platforms or laws regulating online illegal content. Some of the most common types of Borderline content identified in the EU are anti-establishment/anti-institutions, antisemitic, anti-trans, misogynistic, anti-migrants, racist, or against COVID-19 measures.”

³ “Filter bubble” refers to a homogeneous feed/recommended content, caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.

The European Commission's DG CONNECT has also been involved in the development of the study. The study was conducted from July 2022 to May 2023.

The study methodology is novel and discussed in Section 3.2 of the Report. Internal proprietary platform data (including algorithms), data, bots and Application Programming Interface (API) queries, which are common for this type of study, were not used. Trust Lab worked with human analysts to simulate motivated users intentionally seeking TVE content. The simulation was designed to ensure consistency and comparability across all platforms by the use of keyword lists, controlling for the amount of time exposed to each platform and repeating the experiments many times with different agents, thus increasing the likelihood of replicating the study's findings.

To address the six tasks defined by the study's sponsors, the following metrics were used:

- Findability: How easy is it to find harmful content, measured in the count of harmful posts during a one-hour interval?
- Removal Rate: What is the amount/percentage of harmful content removed by the platform?
- Removal Time: How much time did it take for harmful content to be removed by the platform?
- User Sentiment: What is the public (crowdsourced) perception of the appropriateness of harmful content, measured on a 5-point Likert scale?
- Amplification (filter bubble size): What is the amount of Bad Topic (TVE-related content, but not necessarily harmful content), Bad Content (TVE content), and Borderline content in a user's feed, measured as a percentage of the first 30 posts in the feed?

FINDINGS

Trust Lab's analysis found significant amounts of TVE content across all the major social media platforms as well as markets/languages. In addition, evidence of algorithmic amplification of TVE content on those platforms was also confirmed. The study identified significant differences in TVE-related performance and behaviour across all platforms and markets. The study's key findings include the following:

1. Evidence of amplification. The study validates the hypothesis that increased interaction with TVE content and Borderline content results in higher amplification of such content to users. All the platforms showed amplification of TVE-promoting content in their feeds. The percentage of TVE content in each platform's feed did not exceed 10% on average. Across all markets and platforms, the amplification of TVE content increased by 18% and Borderline content by 65%⁴. Twitter showed the highest level of amplification, YouTube ranked the second highest in amplifying TVE content, while TikTok showed the least. Regarding country and languages, platforms show the most TVE amplification occurring for Polish and German and for content related to Left-Wing political affiliation and younger Age Groups.

2. Findability Scores. Twitter had the highest findability scores of TVE content among the five platforms, and YouTube had the lowest findability score. Italian TVE content had the highest

⁴ This is the average percentage change between the first and third (final) amplification measurements across all languages and platforms. See the methodology, Section 3.2, for more details.

findability score of all languages when measured across all platforms. Violent Left-Wing TVE content had a significantly higher Findability score than other TVE types.

3. Removal Rates. More than 90% of the TVE content found by Trust Lab remained on the platforms by the end of the evaluation period 8 weeks later. TikTok removed more TVE content than the other platforms; Twitter and YouTube removed the least. Arabic TVE content was removed more frequently than the other languages, while Italian TVE content was removed less.

4. Borderline Content. Repeated cycles of user interaction and evaluation suggest that the amplification of Borderline content grows over time and at a higher rate than TVE content.

5. Variability across platforms and demographic blindspots. Each platform behaves differently in recommending TVE content with similar user features and behaviours. All platforms tended to recommend TVE at increasing rates as users interacted more with it, regardless of its TVE rating. Left-Wing politically oriented recommendations were far less likely to be removed by platforms than other groups. TikTok had a significantly low rate of success in amplifying TVE content to users, while Twitter and YouTube had the highest, with the former recommending most to Right-Wing and the latter recommending TVE content even to users with low rates of interaction. Ultimately, analysis based on user demographics and platform engagement data is still inconclusive to determine the full scope of how recommender algorithms operate on the back-end; more transparency and research is needed.

TABLE OF CONTENTS

1	OBJECTIVE OF THE FINAL REPORT.....	6
2	INTRODUCTION	6
3	APPROACH.....	7
3.1	THEORETICAL FRAMEWORK	7
3.1.1	Addressing Replicability	7
3.2	METHODOLOGY	11
3.2.1	Contributing Factors to Amplifying Harmfulness.....	12
3.2.2	Process.....	13
3.3	QUALITY.....	14
4	ANALYSIS	15
4.1	TASK 1 - MEASURING THE DISSEMINATION OF TVE CONTENT.....	15
4.1.1	Findability Main Highlights	15
4.1.2	Findability Findings	15
4.1.3	Removal Metrics Main Insights	16
4.1.4	Removal Metrics Findings	17
4.1.5	User Sentiment Main Insights.....	19
4.1.6	User Sentiment Findings.....	19
4.2	TASK 2 - MEASURING THE ROLE AND EFFECTS OF AUTOMATED DISSEMINATION OF TVE CONTENT..	21
4.2.1	Filter Bubble Main Insights	21
4.2.2	Filter Bubble Findings.....	21
4.2.3	Borderline Content	22
4.3	TASK 3 - ASSESS THE RISK POSED BY THE AUTOMATED DISSEMINATION	24
4.3.1	Risk Assessment Main Insights	24
4.3.2	Content Moderation Background	24
4.3.3	Risks Assessment.....	26
4.3.4	Risk Metrics	27
4.3.5	Mitigation Strategies.....	28
4.4	TASK 4 - COMPARE THE PHENOMENON BETWEEN PLATFORMS SHARING THE SAME BUSINESS MODEL	30
4.4.1	Comparison Between Platforms Main Insights.....	30
4.4.2	Platforms	30
4.4.3	Languages.....	31
4.4.4	TVE Type	31
4.4.5	Borderline Content	31
4.5	TASK 5 - ASSESSING THE RISK OF RADICALISATION DUE TO ALGORITHMIC AMPLIFICATION	33
4.5.1	Introduction	33
4.5.2	Analysis of Content Posts	34
4.5.3	Predictive Analyses of Sessions	38
4.5.4	Discussion.....	41
4.6	TASK 6 - PROVIDE GUIDANCE ON CONTENT MODERATION	43
4.6.1	Guidance on Content Moderation Main Insights.....	43
4.6.2	Systems.....	43
4.6.3	Tools	51
4.6.4	Third-party Services	52
4.6.5	Content Moderation Recommendations.....	53

5	CONCLUSIONS AND RECOMMENDATIONS	54
6	APPENDIX.....	56
6.1	PROJECT TEAM	56
6.2	KEYWORDS.....	58
6.3	EXTERNAL EXPERTS	62
6.4	TASK 1: FINDABILITY CHARTS.....	63
6.4.1	Findability per Platform	63
6.4.2	Findability per Language	64
6.4.3	Findability per TVE Type	65
6.4.4	Findability per Language / TVE Type	66
6.5	TASK 1: EXAMPLES OF ITALIAN VIOLENT LEFT-WING EXTREMIST CONTENT	67
6.6	TASK 1: EXAMPLES TWITTER CONTENT	69
6.6.1	Violent Right-Wing Extremism and Borderline Examples.....	69
6.6.2	Violent Left-Wing Extremism and Borderline Examples	72
6.6.3	International Extremism and Borderline Examples.....	74
6.7	TASK 1: USER ENGAGEMENT CHARTS AND P-VALUES	77
6.7.1	Average Number of Shares	77
6.7.2	Average Number of Likes.....	78
6.7.3	Average Number of Comments	78
6.7.4	Average Number of Followers	79
6.8	TASK 1: REMOVAL RATES FOR TVE CONTENT.....	80
6.8.1	Removal Rates per Platform.....	80
6.8.2	Removal Rates per Language	81
6.8.3	Removal Rates per TVE Type	82
6.8.4	Average number of Shares associated with Removed TVE.....	83
6.8.5	Average number of Shares associated with TVE that wasn't Removed	84
6.8.6	Average number of Shares associated with TVE that wasn't Removed broken down by language	85
6.8.7	Average number of Shares associated with Removed Borderline content.....	85
6.8.8	Average number of Shares associated with Borderline content that wasn't Removed	86
6.8.9	Average number of Shares associated with Borderline content that wasn't Removed broken down by language.....	87
6.9	TASK 1: REMOVAL TIME.....	88
6.9.1	Removal Time per Platform	88
6.9.2	Removal Time per Language	88
6.9.3	Removal Time per TVE Type	89
6.10	TASK 1: USER SENTIMENT METRICS.....	90
6.10.1	Severity Ratings by Platform.....	90
6.10.2	Severity Ratings by Language	91
6.10.3	Severity Ratings by TVE Type	92
6.10.4	Severity Ratings over Time per Platform	93
6.11	TASK 2: AMPLIFICATION CHARTS.....	94
6.11.1	Amplification Across all Platforms.....	94
6.11.2	Amplification Average Percent of Bad Content in Feed per Platform	95
6.11.3	Amplification Average Percent of Bad Content in Feed per Language	96
6.11.4	Amplification Average Percent of Bad Content in Feed per TVE Type.....	97
6.11.5	Amplification Percentage Change for Bad Content per Content Type.....	97
6.11.6	Amplification Percent Change for Bad Content per Platform.....	98
6.11.7	Amplification Percent Change for Bad Content per Language.....	99
6.11.8	Amplification per Platform.....	99
6.11.9	Amplification per Language	102
6.11.10	Amplification per TVE Type	106

6.11.11	<i>Amplification P-values</i>	108
6.11.12	<i>Interactivity and Amplification</i>	110
6.11.13	<i>Findability and Amplification</i>	110
6.11.14	<i>Engagement Ratio and Amplification</i>	111
6.11.15	<i>Removal Rate and Amplification</i>	112
6.12	BORDERLINE CHARTS	113
6.12.1	<i>Removal Rates per Platform</i>	113
6.12.2	<i>Removal Rate per Language</i>	114
6.12.3	<i>Removal Rate per TVE Type</i>	115
6.12.4	<i>Removal Time per Platform</i>	116
6.12.5	<i>Removal Time per Language</i>	116
6.12.6	<i>Removal Time per TVE Type</i>	117
6.13	TASK 3: LIST OF REFERENCES	118

1 Objective of the Final Report

The objective of this Final Report is to describe and document the work performed during this study and the outcomes of all completed Tasks. The report is extensively supported by examples from the measurement tasks, with case studies to showcase the role and effects of the algorithmic amplification and other visual aids. The majority of sections are supported by data visualisations in the form of charts, flowcharts, tables and other forms of data representation to present the results and tell the story in the most effective way.

2 Introduction

The existence of Terrorism and Violent Extremism (TVE) content online is one of the most concerning social media trends witnessed over recent years. This trend has been exacerbated by social media algorithms that provide curated experiences to users by recommending content that keeps users engaged with the platform for longer. The severity and amount of TVE is not contained to a few small corners of the internet. It is easily accessible on major online platforms and threatens personal health and safety, peaceful co-existence, and the well-being of all citizens⁵. Social media platforms are not always able to provide a safe online experience, and policies and enforcement practices can vary substantially across platforms.

This project aims to inform the European Union Internet Forum (EUIF)'s work related to the tools and measures used by online platforms to moderate TVE content. Furthermore, this project will provide insights and analysis into the algorithmic amplification of TVE content on five leading social media platforms (Facebook, Twitter, YouTube, TikTok and Instagram) across eight markets/languages (Arabic, English, French, German, Italian, Polish, Spanish and Russian). The overarching research question is how the five social media platforms compare in exposing motivated new users⁶ to TVE content. In order to answer this question, we designed a research method that allows for inter-

⁵ <https://cronfa.swan.ac.uk/Record/cronfa62902>

⁶ The term "New Users" refers to users who are intentionally looking for certain types of content.

platform comparisons and spans a broader range of platforms and languages than has typically been studied in the field of trust and safety research.

Our approach is to critically review the metrics collected relating to how social media platforms' machine learning-based content delivery systems amplify TVE content. The comparison across markets and platforms enables the ranking of individual platforms based on their performance, among other factors, to understand which platform provides the highest exposure and amplification of TVE content. Subject matter experts and industry veterans provide these recommendations to the European Commission based on assessments of the risks posed by the dissemination of TVE content and the potential for radicalisation through algorithmic amplification. Our expertise in trust and safety and online TVE and Borderline content shapes our guidance on content moderation.

3 Approach

3.1 Theoretical Framework

The goal of the study is to simulate the behaviour of motivated new users to research how social media platforms respond to users who perform targeted searches for TVE content. Motivated new users are new users on a given platform who have a targeted interest in TVE content (based on, for example, real-world interaction with TVE concepts) and who are looking specifically for that type of content on the platform.

The use of new accounts provides us with several advantages over using existing accounts. First, setting up accounts and pre-training them on 'benign' content would take a lot of time and resources that the current study scope doesn't allow for. Second, new accounts ensure that we have a more controlled research environment than we would have in the case of pre-existing accounts. In our current approach, we are aware of all the TVE-related content that agents find, and all the content that is recommended in the user feeds is captured and analysed.

If we pre-trained all the accounts on benign content, we would not only have to account for these differences in our analyses, but we would also have a harder time untangling the effects of the recommender system because of the different pre-exposure any of the accounts might have had during this pre-training phase. Therefore, starting with new accounts provides a clean baseline for comparison; even if the business models of the platforms under study are different, all our research on the platform is starting from the same entry point. In order to provide a quantitative analysis, such a common baseline is necessary.

Using new accounts also gives us the opportunity to test how quickly a recommender system will start recommending TVE content to these types of users. This information is relevant because the quicker the recommendation and amplification of TVE content occur, the more implications this has for mitigations on the platform's side. Thus, we are using the 'cold start' problem to our advantage here.

3.1.1 Addressing Replicability

Since user behaviour is at the heart of this study, the concern about the replication of findings must be addressed. In recent years, we have seen that social science research into human behaviour has

faced issues with the replication of findings⁷. This is because human behaviour is complex, and it is unwise and inaccurate to make generalisations from the results of a single study.

Behavioural studies like these will always suffer from issues with replication, at least to a degree. In order to minimise the chance of introducing unwanted variance while at the same time keeping the behaviour under study within the realm of realistic (albeit extreme/targeted), we used the following constraints:

- *Using a consistent understanding of TVE across platforms:* we have developed a set of policy definitions in accordance with the European Commission that covers a spectrum of TVE content and can be used cross-platform, enhancing the focus of our study. A consistent understanding is needed for this study, in part because platform-specific policies are inconsistent when compared to each other.
- These definitions are thus heavily informed by the existing platform definitions for TVE content but go beyond what the platforms currently define as TVE to more adequately capture the extent to which Borderline content exists on these platforms. This means that we expect content might not be removed from the platform if it is not covered by the platform's definition of TVE or that platforms might miss it during enforcement. This provides us with an opportunity to highlight the limitations of the existing platform definitions and provide mitigation strategies to address them.

Definition of TVE: Any media, including text, images, and videos, that promotes or glorifies terrorism or violent extremism or advocates for the use of violence to achieve political, ideological or religious goals.

- This category includes content that is the most explicitly promoting or supporting terrorism or violent extremism and which contains explicit calls to action or direct incitement to violence.
 - The content contains explicit calls to action or statements of support for terrorism or violent extremism.
 - The content contains inflammatory or provocative language that could be perceived as supportive of terrorism or violent extremism.
 - The content contains images or videos that are clearly promoting or supporting terrorism or violent extremism.
- ❖ **Violent Right-Wing Extremism (VRWE)** are acts of individuals or groups who use, incite, threaten with, legitimise or support violence and hatred to further their political or ideological goals, motivated by ideologies based on the rejection of democratic order and values as well as of fundamental rights, and centred on exclusionary nationalism, racism, xenophobia and/or related intolerance.⁸

⁷<https://arstechnica.com/science/2018/08/why-do-only-two-thirds-of-famous-social-science-results-replicate-its-complicated/>

⁸ The definition is provided by the DG HOME and incorporated into our methodology.

- ❖ **Violent Left-Wing Extremism** is a collective term for all efforts directed against the free democratic basic order that is based on treating the values of freedom and (social) equality as absolutes, especially as they are found in anarchist and communist ideas.
- ❖ Regulation Addressing Dissemination of **Terrorist Content Online** (EU 2021/784) (TCO) provides information for EU Member State competent authorities on reporting terrorist content. The full language of TCO can be found in the Official Journal of the European Union⁹. Under TCO, reasons for considering the material to be terrorist content are that such material:
 - 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.

Definition of Borderline content: This category includes 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)
- tactics used to manipulate users and amplify Borderline content leading to violent extremism, such as algorithmic amplification techniques that profit from biases in content sharing algorithms.¹⁰
- ❖ *Having only native speakers performing searches:* the underlying reasoning is that we want to understand language nuances in the platform's approaches to amplifying and/or moderating TVE and Borderline content and to get a better understanding of potential risk areas for specific languages.
- ❖ *Structuring the search tasks when it comes to keywords and search terms:* We used a fixed set of 60 keywords, translated for each language. Keywords were derived from existing research and our own evaluation and internal testing. The 60 keywords we chose are related to common terrorism themes across all markets.

This set is subdivided into three groups of 20 keywords for each type of TVE under study:

- Violent Right-Wing Extremism (Local/National and Ideological)
- Violent Left-Wing Extremism (Local/National and Ideological)
- International TVE (Religiously motivated terrorism)

⁹ Regulation (EU) 2021/784 of the European Parliament and of the European Council of 29 April 2021 on addressing the dissemination of terrorist content online, [2021], L 172/79.

¹⁰ This definition is taken from the EUIF Handbook on Borderline Content (2023).

Personas were assigned to one of three TVE types and only used their set of 20 keywords to provide a balance between being thorough and providing comprehensive coverage. Keywords were chosen according to Trust Lab's proprietary and representative selection criteria. Within this set of 20 keywords, there were a similar number of keywords for people, organisations, hashtags and slogans (or phrases). Agents were able to select keywords from within their set randomly but were not able to add new keywords. All keyword lists were translated by native speakers into terms that are culturally relevant in the local language¹¹. People's names were not translated, but hashtags and slogans may have been changed slightly in order to fit better with a specific language's nuances.

The keywords serve as a basis for the targeted searches. As such, this structure ensures that every search starts at the same place, with very few limitations with regard to how a user's journey subsequently unfolds after the initial search. Thus, we are recreating the 'rabbit hole' experience where a user might start with one video and end up exploring different videos.

The underlying reasoning for this approach is that having the same list of keywords for each language improves comparisons across platforms and languages. Some keywords will unearth more content than others in certain languages or on certain platforms, and being able to report on these differences will add a valuable component to our study. This would not be possible if we were to have different sets of keywords for each language or platform. Having different sets of keywords would also limit our ability to interpret any findings for metrics such as findability. This is because we would not be able to tell whether a certain keyword is a culprit or if a certain platform simply performs better in terms of concealing TVE-related content from its users.

¹¹ The list of English keywords can be found in the Appendix, Section 6.2.

3.2 Methodology

In this field of study, our methodology is comparatively unique. Instead of using bots or automated Application Programming Interface (API) queries, we use human agents (scouts) to simulate 'real' accounts. Their behaviour is designed to ensure consistency and comparability across platforms to enhance the likelihood of replication of the study's findings.

Human agent-driven investigations present certain limitations and challenges that have to be taken into consideration, such as human error, bias, and high costs associated with the method. Yet, we believe the agent-centric approach for data collection is best suited for the task at hand. One reason is that platforms' terms of service do not allow programmatic (automated) searching. Another reason is that the study aims to find out how user behaviour influences the content that is recommended to them (and, consequently, how that content influences the user), and this would not be possible with an automated approach. Lastly, and most importantly, it is the absence of primary data from the platforms themselves that place major constraints on the field of Trust & Safety research. Without access to platform internal data (access to the full corpus on content and behaviours) or a representative stratified sample (also needs to be provided by the platforms due to access restrictions) or a waiver for platform terms of service (TOS) that don't allow for programmatic access and searching, the results will be limited. In our discussions with the platforms, this point was raised multiple times, but broader data access was not provided. As a good sign for future research, the EU's Digital Services Act introduces provisions that allow access to data to researchers of key platforms and requires very large online platforms to disclose key data on the functioning of algorithms, a step very much needed to increase transparency and accountability for user safety.

The current approach is realistic yet constrained. In real life, people may not just use their accounts for searching for TVE content and might switch between different interests and topics, some of which are more innocent than others. A complete 'real life' approach would nevertheless introduce variance of a nature that would make it hard for our study to be replicable by others or to assess our results in a comprehensive manner. Therefore, we have chosen this approach while keeping these trade-offs in mind for future research.

Our study aims to understand how social media platforms' recommender systems algorithmically amplify TVE and Borderline content to users. Recommender systems, at their core, are algorithms that suggest relevant items to users - these could be products to buy, movies to watch, content to consume, and so on. Since there are many different recommender systems on different platforms, this could be an expansive endeavour. Some platforms recommend their users other accounts to follow/subscribe to, some platforms recommend content the user should consume next ('You watched this video, here's another one you might like'), and most platforms have separate 'Explore' pages that are exclusively filled with recommended content.

To ensure comparability across platforms and to keep the scope concise, we have chosen to limit this study to the recommender systems that populate the home feed of users. Feeds exist on all the platforms under study, while other recommender systems can be unique to one or several platforms and would, as such, not make as good a basis for comparison.

To compare platforms and markets directly, results need to be quantifiable, and metrics must be consistent across all platforms. To assess amplification, we have identified the following relevant metrics:

Table 3.2 - Metrics to assess amplification across platforms

Metric	Description
Findability	How many pieces of content a motivated user can find during a targeted search on the platform within a limited time window. This variable is a proxy for the amount of TVE content that is surfaceable on the platform.
Filter Bubble	A homogeneous feed/recommended content caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.
User Sentiment	Crowdsourced severity ratings of the content surfaced. These ratings will be on a Likert scale of 1-5 ¹² .
Removal Rate	How much content that is surfaced is being removed by the platform within a 8-week time period. This is passive removal; agents are not to report pieces of content as this would influence the algorithm.
Removal Time	How long it takes for a piece of content to be removed by the platform within an 8 weeks time period. This will be passive removal; agents are not to report pieces of content as this would influence the algorithm.

Findability is not the same as prevalence. We don't capture the total amount of content that agents come across, nor do we make any claims on how much TVE content actually exists on a platform. Findability is about how much content a person can find within a certain time frame.

Amplification in the context of our study can thus be seen as the increase in the amount of TVE content over time and the relationship between initial user searches and the amount of TVE content being recommended afterwards. Either one of these metrics indicates an amplification of TVE-related content.

In our study, we have been using two metrics ("Bad Content" and "Bad Topic") to measure the depth of filter bubbles. Bad Topic is content that is related to TVE topics but could be neutral, promoting or negative. This includes harmful TVE as well as EDSA¹³ content about TVE. Bad Content is synonymous with TVE content and what is also referred to as promoting TVE. It is also a subset of Bad Topic. We capture this distinction to evaluate whether the recommender systems recommend content to users based on the overall topic of TVE versus the actual content of the posts being consumed.

For the Moderation metrics (Removal Rate and Removal Time), Trust Lab used patented technology "Kaptix", which monitors each found piece of content for removals. Kaptix is an internally developed tool used to monitor whether social media posts are live on the platform.

3.2.1 Contributing Factors to Amplifying Harmfulness

In order to better understand which factors (behavioural or contextual) contribute to amplifying potentially harmful content, we have tracked both content metadata (number of likes, shares, and comments as well as the number of followers of the posting account) and behavioural data from our agents.

¹² A graphic depiction of the values on the Likert scale is available in Figure 4.1.5 of the report.

¹³ Educational, Documentary, Scientific, Artistic.

Agents were divided into High/Low interaction categories. Low Interaction agents only searched and watched content (for videos up to 5 minutes, regardless of video length), while High Interaction agents also liked and followed content. Agents did not comment or re-post/re-share the content. The influence of human interaction and contextual data on the harmfulness risk that recommender systems pose will be discussed in more detail in Task 5.

3.2.2 Process

Agents were assigned one or more personas. Within each language, 30 personas were created. Each persona had an account on each of the five platforms under study. This means that for each language, there were 10 Right-Wing TVE, 10 Left-Wing TVE and 10 International TVE agents, with 5 High Interaction and 5 Low Interaction agents per TVE Type. Agents were instructed to perform three Search Tasks and three Evaluation Tasks for each account. This allowed us to study the changes to the search results and feed recommendations over time.

All tasks were performed on mobile devices.

- **Search Task:** The Search Task started from a seeded keyword search. Agents were either in the Low or High interaction group and interacted with the content as per instructions. Agents were not allowed to add self-made keywords to searches. Agents were allowed also to browse their feeds to search for, interact with and capture TVE content according to the Interaction Type they were assigned to. Each Search Task was limited to one hour. We chose a one hour cutoff to ensure consistency among agents and due to resource constraints.
- **Evaluation Task:** The Evaluation Task started one day after the Search Task. This was to give the recommender systems enough time to update their recommendations. Agents were instructed not to interact with any content but to capture each post in their feed up to the first 30 posts.
- **Repetition:** Both Search and Evaluation Tasks happened 3 times to collect 3 data points for more accurate measurement and monitoring.

Table 3.2.2 - Number of personas per language

TVE-type \ Interaction type	Low Interaction	High Interaction
National / Violent Right-Wing Extremism (20 keywords)	5 personas	5 personas
International Terrorism (20 keywords)	5 personas	5 personas
Other / Violent Left-Wing Extremism (20 keywords)	5 personas	5 personas

With 8 languages and 30 personas per language, that means that we have 240 personas in total.

3.3 Quality

Results from the Search Tasks and Evaluation Tasks were subjected to a Quality Assurance Process to ensure that the content found and captured was indeed TVE content. For this work, we partnered with several international organisations that specialise in finding and analysing Trust and Safety-related content, as well as individuals from academia who study and work in the field of Terrorism and Violent Extremism¹⁴.

Table 3.3 - Quality Assurance

First Level Review	Search: Agent made an assessment of what is and is not TVE-related content. The agents have had policy training to make this distinction. Specialised QA agents reviewed 100% of the content. Evaluation: The agent performs a cursory labelling exercise on the web form. 100% of this sample is reviewed by specialised QA agents.
Second Level Review	A random sample of the complete dataset was used to analyse other metrics (not Findability) to control for the quality of the labelling performed by human raters. Due to the large amount of data that was collected and complexity of the labelling task, multiple rounds of labelling were necessary by qualified individuals on a smaller, cleaner sample. Based on the specifics of the data (amount of search stages, amount of platforms and languages under study), we arrived at a maximum sample of 7200. Since not all data combinations had 30 posts, the total sample size was 7074. The data was ensured to have the following: - Is the post related to TVE? (binary, based on average of rater's scale) - Is the post promoting TVE? (binary, based on average of rater's scale) - Is the post Borderline TVE content? (binary, based on average of rater's scale) - Content Appropriateness score (scale, crowdsourced) - Engagement Labelling (sourced from post data) To calculate the Findability scores, we used the complete set of content that was found during the Search phase. We relied on the first-level review to determine whether or not a piece of content was TVE.
Third Level Review	Content from the original dataset that had already been reviewed by Trust Lab experts and third-party academia experts was also included in the stratified sample as described above.

¹⁴ More details on the consulted experts can be found in the Appendix, Section 6.3.

4 Analysis

Due to the large volume of content we captured through our work we have chosen a stratified sampling method for the analysis of our Filter Bubble metrics and Removal Metrics. Findability scores were calculated using the complete set of data that was found during the Search Stages rather than a sample. User Sentiment data was also calculated based on the stratified sample.

4.1 Task 1 - Measuring the Dissemination of TVE Content

4.1.1 Findability Main Highlights

- **Twitter has the highest Findability score of all platforms analysed.** Since we don't know how specific characteristics of platforms (such as having an open API yes/no) influence moderation of content, we do not make assumptions on the potential impact of these characteristics on the amount of content we were able to find.
- **Arabic TVE has the lowest Findability score, while Italian has the highest.** Finding TVE content through search appears to be easier in Italian, while it's harder in Arabic. There are a lot of confounding factors that could contribute to this finding, so we cannot make assumptions about the cause.
- **Left-Wing TVE has a higher Findability score than other TVE Types.** This points towards a larger trend, where International and Right-Wing TVE content are more easily identifiable as extreme content, whereas Left-Wing TVE constitutes more variations across cultures and languages.

4.1.2 Findability Findings

Our analysis shows that Twitter has the highest Findability score (2.91) on the platform¹⁵, while YouTube has the lowest (1.71) among the platforms we analysed. This means that Facebook and Instagram, as well as TikTok, make up the middle on our scale. Findings for YouTube and Twitter are statistically significant. (See Figure A.6.4.1).

Moving to languages, Arabic TVE has the lowest Findability score (1.23), while TVE in Italian language has the highest (2.78). This means that finding TVE content through search is easier in the Italian language, while it's harder in Arabic. Both of these findings are statistically significant. (See Figure A.6.4.2).

For TVE Types, International TVE has the lowest Findability score (1.41), while Left-Wing TVE has the highest (3.01). That means that it's easier to find Left-Wing TVE, while it's harder to find International TVE by searching the platforms. These findings are statistically significant. (See Figure A.6.4.3).

Deep Dive: Violent Left-Wing Extremist Content

Since Violent Left-Wing Extremist content has a significantly higher Findability score than other TVE types, we wanted to explore this category of content in a bit more depth. The kind of content we find

¹⁵ Examples can be found in the Appendix B, Section 6.6.

when we look for Violent Left-Wing Extremist content is mostly supportive of anti-capitalism, pro-anarchism and pro-communism¹⁶.

These findings point towards a larger trend. Terrorist and Violent Right-Wing Extremist (TVRWE) topics, such as neo-Nazism, anti-Semitism, racism and white supremacy, frequently attract the attention of policymakers and journalists, which causes social media platforms to invest in technical and human resources to reduce the harm of such material. On the other hand, Violent Left-Wing Extremist content encompasses a range of topics that have entered modern political, social and cultural spheres without causing the same level of concern, such as anti-capitalism, general statements about the failure of modern-day institutions and the desire to learn about other forms of governance, such as communism and anarchism.

Deep Dive: User Engagement Metrics

In order to better understand how many people might have engaged with the TVE content that we surfaced, we performed an engagement analysis on four metrics: likes, comments, shares and the follower count of the accounts who posted the TVE content.

- TVE content on Instagram and YouTube is shared much less widely than content on other platforms. Content on Twitter was shared more widely. (See Figure A.6.7.1)
- TVE content on YouTube is liked more on average than other platforms. Facebook has the lowest number of likes for TVE content. (See Figure A.6.7.2)
- TVE content on YouTube is commented on more on average than other platforms. Comments on Instagram seem virtually non-existent. (See Figure A.6.7.3)
- TVE content on YouTube comes from accounts with more followers on average than on any other platform examined. (See Figure A.6.7.4)
- While YouTube has a low Findability score, the platform scores high on almost all engagement metrics, except for shares. This seems to indicate that the platform is in control of suppressing Bad Content from being found in search but that users have found other ways to access the content. (See Appendix, Section 6.7 for more details)
- Facebook and Instagram are generally at the bottom of the engagement charts. Especially Instagram scores low, having no shares and no comments on average for TVE content. (See Section A.6.7)
- Comparing the TVE vs non-TVE engagement averages shows that TVE content gets much lower engagement than non-TVE content, but these results are not statistically significant for most platforms. (See Appendix, Section A.6.7 for more details)
- It should be noted that the differences in engagement between platforms could also be due to the user interfaces of the platforms. It could be argued that YouTube and Instagram are less set up for sharing content than Twitter or TikTok, for instance. (See Appendix, Section 6.7 for more details).

4.1.3 Removal Metrics Main Insights

- **The platforms removed very few items** during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it, resulting in TVE being shared heavily on some platforms like Facebook.

¹⁶ Examples can be found in the Appendix, Section 6.6.2.

- **Twitter removes the least amount of TVE content.**
- **Arabic language TVE content is removed more often than other types of TVE content, while Italian TVE content is removed less.**
- **Terrorist and Violent Left-Wing Extremist content is also removed less than other types of TVE content.**
- Findability and Removal Rate are related to each other: the less likely it is that TVE content is removed, the more likely it is that we can find it. **It is, therefore, not surprising that the findings in this section resemble the findings from the Findability section.** Twitter removes the least amount of TVE content, and Arabic TVE content is actioned and removed more, while Italian TVE content is removed less, and Left-Wing TVE content is also removed less, than other types of TVE content.
- **Combining these insights seems to indicate that platforms tend to dedicate less effort to removing terrorist and Violent Left-Wing Extremist content and Violent Extremist and terrorist content in Italian language and instead focus more on removing violent Extremist and terrorist content in Arabic language, International and Right-Wing TVE content.** Since platforms still remove less than a third of the content marked as TVE Promoting, this seems to point towards inconsistent content moderation policies, in particular, less pronounced policies and enforcement focus towards markets and topics that garner less attention from the general public.
- **YouTube scores low on Removal Rate, which is an interesting reversal case:** The platform does comparatively well with suppressing TVE content in search without removing such content from the platform. See Section 4.2.2 for a deep dive into this.

4.1.4 Removal Metrics Findings

Our analysis shows that TikTok removed more TVE content than other platforms (11% of the content removed), while YouTube removed the least amount of content (3% of the content removed). Overall, platforms removed less than a third of the content marked as TVE promoting. These findings are statistically significant. (See Figure A.6.8.1).

Arabic language TVE content is more often removed than most other languages (10% of the content removed), while Italian language TVE content is less often removed than most other languages (3% of the content removed). These findings are statistically significant. (See Figure A.6.8.2).

TVLW content is less often removed than other TVE types: only 4% of the time, compared to 7% (Right-Wing TVE) and 8% (International TVE). This finding is statistically significant. (See Figure A.6.8.3).

The platforms removed very few items during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it. All Removal Time metrics are anecdotal because of the small number of removals overall. At each interval, the

percentage represents the amount of content that was removed out of the total removed within an 8-week period. (See Figure A.6.9.1).

Generally, German language TVE content was removed more quickly, while Italian TVE content was removed more slowly than other languages we've studied. (See Figure A.6.9.2) TVLW content also seemed to be removed the slowest, while International TVE content seemed to be removed more quickly. (See Figure A.6.9.3).

Deep Dive: Removal Metrics

The low rates of removal by themselves do not tell a very satisfactory story, so a deeper analysis was conducted to look into the impact that this is causing based on the data that was collected. The focus of this analysis was to contrast removals with user engagement and to review in particular, the number of shares. Shares were chosen because this is a much more deliberate action that also highly impacts the dissemination of TVE content than other forms of engagement¹⁷. It's worth noting that shares are most likely the least accurate for platforms like YouTube that have a significant presence on desktops because sharing can be just as easily done by copying the URL from the address bar. Also, sharing off-platform (for example, sharing a link on another social network) may not be trackable by the content owner.

Of the content that was eventually removed from platforms within the study period, TikTok TVE content was disseminated the most, with 120 shares per TVE post on average (see Figure A.6.8.4). However, of the TVE that wasn't removed by platforms Facebook has the highest number of shares per post 341 followed by TikTok with 283 on average (see Figure A.6.8.5).

The proportionally higher removal rates of TikTok contrasted with higher average number of shares suggests that TikTok TVE content is being consumed and spread at higher rates than other platforms. Thus, despite TikTok's higher performance on other metrics, it is struggling to keep up with its user base and high throughput even on high-harm content like TVE. This same trend appears in Borderline content except with higher shares counts (see Figure A.6.8.7).

Looking at how different languages impact TVE content that wasn't removed, French and Italian have more shares on Facebook while other languages are more prevalent on TikTok (see Figure A.6.8.6). Borderline content in German language, however, sees more shares on Facebook which is the only language that changes between TVE and Borderline content (see Figure A.6.8.9).

¹⁷ Ljungberg, J. et al, Like, Share and Follow: A Conceptualisation of Social Buttons on the Web, July 2017.

4.1.5 User Sentiment Main Insights

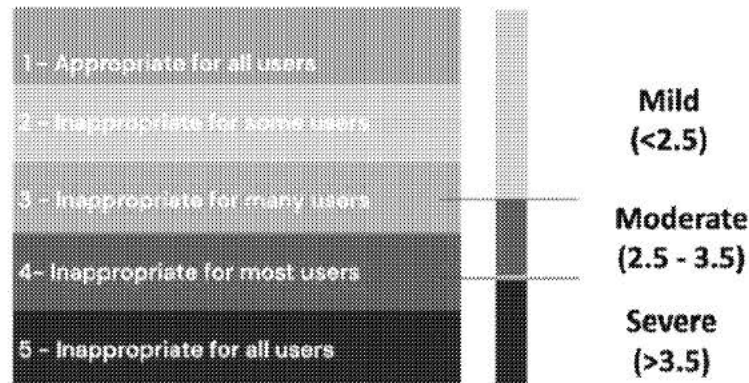


Figure 4.1.5 - Appropriateness score

- **Platforms are mostly similar** in terms of **severity of content based on user ratings**, indicating that severe content is spread roughly equally across the platforms. Italian content is rated as the most severe, while German is rated as the least severe. Most differences between languages and their severity ratings are statistically significant.
- **TVE content in Italian is rated as the most severe, continuing a trend that seems to indicate that Italian language TVE content is largely going unnoticed by platforms.** Other languages are also mostly statistically significantly different from each other, indicating that users of different languages have potential cultural differences when it comes to how harmful they judge TVE content to be.

4.1.6 User Sentiment Findings

Platforms are mostly similar in terms of how severe users rate their content, indicating that severe content is spread roughly equally across the platforms. (See Figure A.6.10.1).

Italian language TVE content is rated as the most severe (98% rated as severe), while German language TVE content is rated as least severe (43% rated as severe). Most differences between languages and their severity ratings are statistically significant. (See Figure A.6.10.2).

Right-Wing TVE content is viewed as less severe by users than other TVE types (78% rated as severe), which is a statistically significant finding. This is interesting and could be due to the fact that users are getting more used to seeing Right-Wing TVE content, but also to the fact that platforms are removing the more severe Right-Wing TVE content from their platforms more readily. (See Figure A.6.10.3).

Severity Over Time

Does searching for and interacting with TVE content influence the severity of the content that is subsequently found and recommended to users? A trend here could be an indication of amplification, not necessarily in the *amount* of content that a user sees, but in the *harmfulness risk* of the content that users engage with.

No clear trends emerge when looking at the severity ratings over time per platform. (See Figure A.6.10.4).

4.2 Task 2 - Measuring the Role and Effects of Automated Dissemination of TVE Content

4.2.1 Filter Bubble Main Insights

- **Taking all platforms, languages and TVE types together, there is an amplification of all content types over time.** This amplification is only significant for Bad Topic and Borderline content, but not for Bad Content.
- **All platforms show signs of amplification after the first Search phase, showing Bad Content in the user's home feeds, but platforms do not show any significant further amplification of Bad Content over time.** On average, no more than 10% of any feed consisted of Bad Content.
- **All languages show signs of amplification after the first Search phase, but only Italian shows significant further amplification of Bad Content over time.** French does show a significant further amplification of Borderline and Bad Topic content over time.
- **All TVE Types show signs of amplification after the first Search phase, but none show significant further amplification of Bad Content over time.**
- Overall, these findings seem to indicate that although there seems to be an amplification of Bad Content in the feed after showing initial interest by a user, this is not significantly further amplified when the user continues to engage with said content.

4.2.2 Filter Bubble Findings

Taking all platforms, languages and TVE Types together, there is an amplification of all content types over time. For Bad Content, the main focus of our study, our amplification metric increased from 4.4% to 5.2%, an increase of 17.5%. Looking at Bad Topic, which includes TVE and non-harmful content related to TVE topics, we saw an increase from 6.9% to 9.6%, which is statistically significant (+39.1%). Additionally, we looked at Borderline Content, which amplified from 3.3% to 5.4% (+65.1%), which is also statistically significant. (See Figures A.6.11.1).

Instagram is the only platform that shows a significant amplification across the three stages for Bad Topic content (+118%), as well as Borderline content (+150%). None of the platforms show a significant further amount of amplification for Bad Content, although Facebook doubled the amount of recommended Bad Content between the First and Third stage. TikTok showed no increase, and YouTube had a negative amplification (-23%); all not statistically significant. (See Figures A. 6.11.8a-e).

Italian and French are the only languages that show significant amplification of TVE related content over time. Italian is significant for Bad Content only (+365%). The French amplification is only significant for Bad Topic (+293%) and Borderline content (+685%), not for Bad Content. Polish has the highest average percentage of Bad Content in the feed (11%), Arabic has the lowest (2%). (See Figures A.6.11.9a-h).

None of the TVE types show significant amplification over time. (See Figures A.6.11.10a-c).

Higher interaction with TVE content does result in a higher initial percentage of Bad Content in the feed (8% for High Interaction, 1% for Low Interaction). (See Figure A.6.11.12).

Deep Dive: How Platforms Reduce TVE Content

When looking across the Findability and Filter Bubble metrics, we can compare how platforms choose to reduce the spread of TVE content. (see Figure A.6.11.13).

- **YouTube:** has the lowest Findability score but a high average percentage of Bad Content in the feed.
- **TikTok:** reflects the opposite trend, having a moderate Findability score and the lowest average percentage of Bad Content in the feed.
- **Twitter:** has the highest Findability score and also the highest average percentage of Bad Content in the feed.
- **Instagram:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.
- **Facebook:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.

YouTube seems to have tighter filtering of search results, while TikTok has tighter filtering of the recommendations in the feed. Relatively speaking, Twitter has the most on both dimensions. Facebook and Instagram appear to have moderate amounts of filtering in both search and feed.

These findings could be the result of YouTube's and Twitter's algorithms having stronger personalisation, combined with a lack of filtering. When platforms strongly personalise content based on past activity but lack adequate filtering of Bad Content, the filter bubble effect becomes most problematic. TikTok's algorithm is known to have strong personalisation effects¹⁸, but seems to have better filtering of TVE-related content than YouTube and Twitter.

4.2.3 Borderline Content

Special interest from the European Commission has been expressed with regard to the effect of the spread of certain types of Borderline content, such as disinformation and some forms of hate speech that can lead to radicalisation. This section will shine more light on how Borderline content plays into the analysis of the metrics that we calculated. It is content that does not meet the thresholds to be labelled as terrorist and violent extremist content, but that can still lead to violent extremism and radicalisation pathways.

Amplification of Borderline Content

- Overall, Borderline content has the lowest initial amplification of all TVE-related content types, but the percentage of Borderline content that is recommended to users has the highest amplification over time of all TVE-related content types.
- Instagram and Twitter are the only individual platforms that show a significant amplification in the amount of Borderline content that is recommended to users over time. Instagram shows 3 times more Borderline content when comparing the First to the Third phase.
- French is the only language that shows a significant amplification of the amount of Borderline content being recommended to users over time. 8 times more content is recommended to users in the Third phase compared to the First phase.
- None of the TVE Types shows a significant amount of amplification for Borderline content.

¹⁸ <https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity>

(See Figures A.6.11.8a-A.6.11.10c).

Removal of Borderline Content

- Overall, all platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any recommended content at all. Since we are comparing all platforms based on the same definition of Borderline content, this gives us a clear picture of which platforms are allowing more Borderline content on their feeds.
- Arabic Borderline content is removed more often during Search than most other platforms. None of the values for Evaluation are significant.
- Borderline content that could lead to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant.
- Removal Times for all TVE content and Borderline content are largely similar for all dimensions.

(See Section 6.12 for figures).

4.3 Task 3 - Assess the Risk posed by the Automated Dissemination

4.3.1 Risk Assessment Main Insights

- **Amplification: What is the prevalence of TVE content in the platform's feed?** Twitter was the platform with the highest level of amplification of Bad Content to the user, followed by YouTube, Instagram and Facebook. TikTok had the lowest level of TVE content amplification or harmfulness.
- **Risk Ratings.** We assigned a composite risk rating for each platform, based on an evaluation of three essential components: Findability, Amplification and User Sentiment. The risk ratings for the platforms mostly followed the percentage of TVE content found in the feed because the feed is where the bulk of the TVE content consumption likely occurs. The risk rating was consistent with removal and engagement rates, and there were fairly big differences among the platforms on this risk metric. Twitter had the highest risk rating of the platforms; TikTok had the lowest risk rating. Facebook, Instagram and YouTube had medium risk ratings.
- The average percentage of TVE content in the feed did not exceed 10%, even for motivated users, and did not seem to increase above 10%, even under repeated user interactions with TVE content. If platforms were only focusing on personalisation, this percentage would likely have been higher. This suggests that platforms either do not personalise much in general or that platforms limit personalisation, when TVE content, topic or Borderline content is detected. That said, when we compared the platforms on different dimensions, we observed clear differences among the platforms and further room for improvement, particularly for platforms such as Twitter with higher risk.

Recommender systems (RSs) are increasingly being used to support decision-making and improve user experiences across a wide range of industries and applications. However, it is important to acknowledge and effectively manage both technical and non-technical risks in order to reap their benefits and promote user safety fully. As requested by the European Commission, this report provides a comprehensive overview of recommender systems and the technical and non-technical risks associated with recommender system algorithms and outlines a range of mitigation strategies to help reduce risks and ensure best practices for the recommender systems and those who use them.

4.3.2 Content Moderation Background

Content moderation, or the process of monitoring, reviewing, and controlling user-generated content on online platforms to ensure compliance with community guidelines, terms of service, or content policies, is a fundamental part of the internet's infrastructure and success. Content moderation involves identifying and removing inappropriate, harmful, or offensive content, as well as addressing violations of rules and regulations set by the platform or applicable laws. The way a platform is designed and how users interact with it significantly influence the posting and interaction of user-generated content, consequently impacting how companies perform content moderation on their services (Grimmelman, 2015).

As discussed, platforms which rely on user-generated content often utilise recommender mechanics to deliver personalised content to users, deliberately surfacing tailored content most likely to keep the user engaged (O'Callaghan et al., 2014). Recommender systems have the potential to inadvertently recommend terrorist or extremist content to users who are not actively seeking such

material. This occurs because the algorithms prioritise user engagement without parallel automated objectionable content detection, allowing such terrorist or extremist content to be surfaced based on predicted engagement levels. Recent studies indicate that inadvertently exposing users to extremist content based on their expressed interest can lead to a reinforcement of extremist content in their feed; a consistent consumption of extremist content may contribute to a shift in their worldview (Edwards & Gribbon, 2013). Platforms face a challenging task in distinguishing between intentional bad actors and users who unintentionally engage in abusive behaviour, adding complexity to the moderation ecosystem. The following breakdown of prevalent filtering mechanisms and technological content moderation systems highlights the intricate balance between preserving the rights of users to freedom of expression and the challenging task of detecting and removing content in a scalable manner.

Content-Based Recommendation Systems

Content-based filtering is a technique used by online platforms, including user-generated content platforms like YouTube, to analyse a user's preferences and create a content engagement profile based on keywords and tags. This approach considers factors such as titles, descriptions, and tags to determine similarities and relevance when recommending videos, particularly for new or unseen content (Zhang, Lu & Jin, 2021). While content-based filtering offers scalability and independence from other users' data, it has limitations in the type of content it suggests to users. In the context of moderating terrorist content, platforms that rely heavily on content-based filtering pose challenges in effectively identifying and addressing such material. The focus on matching user preferences may overlook objectionable content that falls within the grey area of extremism. As the content becomes more extreme, the videos the system relies on for tags may begin to overlap between less extreme and more extreme content, making it less likely for a target audience to report such material to the platform. This increases the risks associated with content moderation and the potential for terrorist or extremist content to go undetected or unaddressed.

Although platforms basically work with some combination of recommender and flagging subsystems, these can be brought together in different ways, sometimes without a clear demarcation between the two. Also, while these subsystems draw upon the basic recommendation algorithms and flagging approaches outlined here to optimise for the metrics specified in section 4.4.2, there are many variations of these algorithms and metrics. Also, the practical pressures of running a business (e.g., generating revenue and retaining users, keeping infrastructure and machine costs low as well as addressing user and customer complaints and escalations) implies that these systems have accumulated numerous optimisations as well as business rules, making them exceedingly complex. As a real-world example, please refer to the recently outsourced Twitter recommender system.

Automated Flagging System

Automated detection systems utilise various techniques such as machine learning algorithms, near-neighbour algorithms, fuzzy hashing models, and MD-5 hashing models to identify harmful or platform-violating content, including both well-known and unknown terrorist and extremist material. These systems bring potentially illegal or policy-violating content to the attention of a team of content moderators, either directly employed by the platform or affiliated with a vendor company. Depending on the specific model employed, there are cases where reports are automatically closed if signals suggest the content is spam, does not violate the platform's Terms of Service, or has been previously reviewed by a content moderator. Automated detection systems play a crucial role in helping online platforms identify more severe content, initiating the cycle of moderation and policy enforcement on the platform.

User-Based Flagging System

User-based reporting systems rely on users to flag content that they believe violates an online platform's policies or is deemed objectionable. These reporting systems vary across platforms, often employing dedicated reporting forms or allowing users to directly flag content they find problematic while viewing it. User flags serve as a means for users to report content that may not have been detected by technical solutions, giving them a voice in the content moderation process. After users submit reports, the platform's team of content moderators reviews them and determines the appropriate actions to take regarding the reported content or offending user account. Many large online platforms deprioritise these user reports in favour of automated detection, but will no longer be allowed to do so under the new European Union's Digital Services Act, which places emphasis on users' right to report and appeal content they find objectionable, including terrorist and extremist content.

4.3.3 Risks Assessment

How can platform algorithms cause filter bubbles?

The term "filter bubble" is generally credited to internet activist Eli Pariser circa 2010. In Pariser's influential book, *The Filter Bubble* (2011), it was predicted that individualised personalisation by algorithmic filtering would lead to intellectual isolation and social fragmentation. To increase user engagement, social media companies may connect users with ideas they are already likely to agree with, thus creating echo chambers of users with very similar beliefs. The concern is that recommender systems may influence users to engage in progressively narrower content domains and in directions they might not otherwise have pursued. The EU Counter-Terrorism Coordinator has argued that the amplification of legal but harmful content may be conducive to radicalisation and violence because it normalises it and exacerbates polarisation in society.

The Global Partnership on Artificial Intelligence (GPAI) has also identified that recommender systems can amplify user bias related to TVE content, as follows:

The key issue for recommender systems . . . is that social media users are known to show small biases towards extreme content of various kinds, that act as another influence on the content items they engage with. For instance, they have a tendency to share political messages containing 'moral emotional expressions' (Brady et al., 2017; Brady and van Bavel, 2021), particularly negative ones (Crockett, 2017; Brady and van Bavel, 2021), messages that refer to a political 'out-group' (Rajthe et al., 2021), and messages that contain falsehoods (Vosoughi et al., 2018). If these biases persist while a user interacts with a recommender system, the system's repeated updates of its user model may lead the user towards messages containing increasing levels of negative political emotions, an increasing focus on political out-groups, and increasing amounts of misinformation – and potentially towards domains of violent extremism. Again, our earlier report (GPAI, 2021) presents these concerns and the studies that support them in detail.¹⁹

¹⁹ GPAI 2022, *Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions*. Report, November 2022. Global Partnership on AI.

Recently, the Global Internet Forum to Counter Terrorism commissioned a survey of the existing empirical literature on the issue of whether recommender systems promote extremist content and whether online users are “radicalising by algorithm”.²⁰ In the literature review, 10 of 15 studies demonstrated that recommender systems can promote extreme content.

Can bad actors exploit platform algorithms?

Terrorists could “game” flagging systems by experimenting and learning over time how to have their content bypass or circumvent either automated or manual flagging systems. According to certain researchers, the survival of certain TVE content on platforms relates to the way in which terrorist organisations have learned to modify their content to evade platform controls. Tactics include: breaking up text and using strange punctuation to evade platform tools that would search for keywords; blurring the terrorist organisation’s branding or adding the platform’s own video effects; mixing the terrorist organisation’s material with content from real news outlets; adding the branding of mainstream news outlets over the top of the terrorist organisation’s content; hijacking platform accounts; and posting tutorial videos to teach other terrorists how to do the same.²¹

Terrorists could also employ fake engagement and fake profiles to promote their content, and if undetected by platforms, recommender systems could unwittingly further promote such content. From gaming of views²² to crowdturfing and astroturfing²³, there are a variety of technical approaches to gain (perceived) popularity on social media, which are already known to be leveraged by foreign actors to influence operations. For instance, in one study, the BBC has reported that a network of Facebook groups was used in an attempt to change perceptions related to the ongoing war in Ukraine.²⁴

In this section, we presented many possible risks that can cause the promotion of TVE content on platforms. While we do not have access to internal platform documentation to understand how each individual platform’s algorithms and moderation systems work, we can assess risk based on externally observable metrics from this study. We do this in the next section.

4.3.4 Risk Metrics

The previous sections described the various risks that recommender systems present, but without access to each platform’s internal, proprietary algorithms and data, we were unable to quantify platform risk directly based on these factors. Instead, we evaluated the overall risk of each platform based on the data collected for this project, which could be observed outside-in. We considered three factors:

²⁰ Whittaker, J., Recommendation Systems and Extremism: What Do We Know? – Global Network on Extremism and Technology, Insights, 17 August 2022. <https://gent-research.org/category/insights/>.

²¹ Corera, G., *ISIS ‘still evading detection on Facebook’, report says*, BBC News, 13 July 2020. See also Nimmo, B. and Hutchins, E., *Phase-based Tactical Analysis of Online Operations (The Online Operations Kill Chain: A model to analyze, describe, compare, and disrupt threat activity from influence operations to cybercrime)*, Carnegie Endowment for International Peace, March 16, 2023.

²² The Flourishing Business of Fake YouTube Views. <https://www.nytimes.com/interactive/2018/08/11/technology/youtube-fake-view-sellers.html>

²³ <https://en.wikipedia.org/wiki/Astroturfing>

²⁴ Putin’s mysterious Facebook ‘superfans’ on a mission. <https://www.bbc.com/news/blogs-trending-61012398>

- User Sentiment: This is a proxy for the severity of TVE content found
- Findability: This is a proxy for the amount of TVE content that is surfaceable on the platforms
- Amplification: This is a proxy for the prevalence of TVE content in the feed.

We saw relatively minor differences in User Sentiment (see Figure A.6.10.1) among platforms. The platform with the highest level of Findability of TVE content (see Figure A.6.4.1) was Twitter, followed by Facebook, Instagram, and TikTok. The platform with the lowest level of Findability was YouTube.

Risk Outcome

We have assigned a composite risk rating for each platform based on our evaluation of three component metrics: User Sentiment, Findability, and Amplification. The risk ratings below incorporate information from the TVE content found in the feed and search and is consistent with removal and engagement rates.

Table 4.4.4 - Risk Rating per Platforms

Twitter	Youtube	Instagram	Facebook	TikTok
Highest	Medium	Medium	Medium	Lowest

The data we collected is based on simulating motivated users looking for TVE content and does not represent the experience of an average user of the platform. We also did not have access to internal proprietary platform data, which would have been needed for comprehensive measurement. That said, as previously discussed in the introductory section, it made sense to simulate and study the motivated user scenario. Also, the consistency observed for different metrics, such as feed prevalence, removal rates and engagement rates, suggests that the data is not an outlier.

The next section presents possible mitigation measures to consider, both based on observed data as well as conceptual risk factors discussed in previous sections.

4.3.5 Mitigation Strategies

There are a few different ways in which platforms can counter filter bubbles:

- By effectively detecting terrorist and extremist content, platforms can proactively filter such content from user feeds and search results. The **Global Internet Forum to Counter Terrorism (GIFCT)** is an Internet industry initiative to share proprietary information and technology for automated content moderation. As noted on the [GIFCT Wikipedia page](#), GIFCT maintains a database of perceptual hashes of terrorism-related videos and images that are submitted by its members and which other members can voluntarily use to block the same material on their platforms. The material indexed includes images, videos and will be expanded to include URLs and textual data such as manifestos and other documents.

- Filter bubbles and echo chambers can be countered by having recommender systems explicitly optimise for content or opinion diversity. There is evidence suggesting that personalisation algorithms can enable exposure to “ideologically cross-cutting content” from outside self-selected echo chambers. (Flaxman et al., 2016). It has also been suggested that they can also be leveraged in improving the automated detection of radicalisation and for facilitating targeted counter-messaging interventions to combat radicalisation. (Schmitt et al., 2018).²⁵ Recent literature has also suggested that algorithms can be the cure for algorithmic filter bubbles.²⁶
- Platforms should ensure that the algorithms themselves don’t have ideological bias. A recent Twitter study (“the most comprehensive audit of an algorithmic recommender system and its effects on political content”) indicated that the political right enjoys higher amplification compared to the political left. Our own data presented in previous sections also examines differences between Right- and Left-Wing TVE content both in terms of discoverability and moderation. Armed with such data, platforms can take a deeper look at their own algorithms and implement necessary mitigations.

²⁵ Wolfowicz, M., Weisburd, D. and Hasisi, B. (2021), Examining the Interactive Effects of the Filter Bubble and the Echo Chamber on Radicalization, *Journal of Experimental Criminology* (2023) 19:119-141 at 136. <https://doi.org/10.1007/s11292-021-09471-0>

²⁶ Gao, C. et al., Counterfactual Interactive Recommender System (CIRS): Bursting Filter Bubbles by Counterfactual Interactive Recommender System, *Computing Research Repository (CoRR)* in arXiv, abs/2204.01266 (2022).

4.4 Task 4 – Compare the Phenomenon Between Platforms Sharing the Same Business Model

4.4.1 Comparison Between Platforms Main Insights

- Among the platforms, Twitter surfaced and recommended the most TVE content to its users and performed below other platforms in terms of TVE content removal.
- Although YouTube's Findability score was the lowest among the platforms, its removal metrics suggested it has room to improve on removing TVE content that exists on the platform. YouTube generally scored higher than the other platforms on user engagement (e.g., likes, shares, comments, number of followers) with TVE content. This seems to indicate that YouTube was effective in suppressing TVE content from users in search but still enabled users to find TVE content through other means and interactions.
- TikTok performed moderately when it came to the Findability of TVE content, but of all the platforms, it recommended the least amount of TVE content to users. This suggests that TikTok may filter its feed more tightly than its search, resulting in no significant amplification of TVE content in the feed. TikTok was generally slow to remove TVE content; TikTok took more than four weeks to remove the content we identified as TVE. TVE content that the platform didn't remove was shared twice as much as TVE content that the platform did remove.

4.4.2 Platforms

Of the platforms, Twitter surfaced and recommended the most TVE content to its users. Twitter was lower than other platforms when it came to TVE content removal. We note that this study was conducted during a period when significant governance, policy and safety process changes were occurring at Twitter. As such, Twitter's results may be significantly different compared to earlier in the year.

YouTube has come under scrutiny during other research projects²⁷ and has since made improvements. Although YouTube's Findability score is the lowest among the assessed platforms, its Removal metrics combined with YouTube's engagement metrics for TVE content suggest the platform still has room to improve on removing TVE content that already exists on the platform.

In our study, YouTube recommended the most TVE content to its users; Twitter was second. It should be noted, however, that all the platforms had weak Removal rating scores, removing 10% or less of the content we marked as TVE-related.

YouTube generally scored higher than the other platforms on user engagement with TVE content. This seems to indicate that YouTube is effective in suppressing TVE content from users in search. But, it still enables users to find TVE content through other means and interactions.

²⁷<https://www.technologyreview.com/2020/01/29/276000/a-study-of-youtube-comments-shows-how-its-turning-people-onto-the-alt-right/>

<https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content>

4.4.3 Languages

Italian TVE content is an interesting case with regard to many of the metrics considered. The Findability score for Italian TVE content was one of the highest, and removal metric scores were in the lowest category. Under the study, Italian users in the sample were also the ones who rated their TVE content as the most severe of all the languages. This suggests that Italy may not be a high-priority market for platforms in terms of Italian-specific automated detection methods, Italian moderators, or other measures integral to a Trust and Safety enterprise.

TVE content in Arabic has long generated special attention from the platforms' moderation teams. Its Findability score is the lowest. Significantly, Arabic TVE content, especially International Arabic TVE content, is more often removed than content from other languages. Since the Paris attacks in 2015, there has been more focus from policymakers and media outlets on social media platforms to police the spread of International TVE from Arabic-speaking countries and regions, which could explain this finding.

4.4.4 TVE Type

Left-Wing TVE content treatment stands out. It's easier to find Left-Wing TVE content on the different platforms, especially in the Italian language. Left-Wing TVE content is also the least often removed. Apparently, this type of TVE content is also not as readily on the radar of platforms, possibly because Left-Wing TVE content covers a broad spectrum of content that is not TVE related per se.

This can be contrasted with Right-Wing and International TVE content, which receives more attention from policymakers and the media, thus incentivising social media platforms to create policies to moderate this type of content more readily.

Overall, one conclusion may be that the public attention to certain types of TVE content, often related to current events, plays a large role in what content gets moderated on platforms. If this is true, one could argue that this is an insufficient way of moderating TVE-related social media content. The platforms could certainly do more to be more consistent in their moderation practices (see also Task 3).

4.4.5 Borderline Content

When we look across the entire dataset, Borderline content appears to be amplified the most over time out of all the TVE-related content types that we've studied, even though Borderline content has the lowest amount of initial amplification.

Zooming in on different dimensions, such as platform, language, or TVE type, this trend is less pronounced. For example, Borderline content does not appear to be significantly amplified over time on any platform except for Instagram and Twitter. French is the only language where Borderline content was significantly amplified over time. Borderline content, potentially leading to terrorist and Violent Right-Wing Extremism, was not significantly amplified over time.

Platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any

recommended Borderline content. Borderline content in the Arabic language is removed more often during Search than other platforms, but none of the values for Evaluation are significant.

Borderline content potentially leading to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant. These findings are similar to other TVE-related content, which seems to show that the policies of the platforms do not place special focus on Borderline content.

4.5 Task 5 - Assessing the Risk of Radicalisation due to Algorithmic Amplification

Assess the level of risk of radicalisation to violent extremist ideologies and terrorism linked to the voluntary or involuntary automated dissemination of terrorist and violent extremist content through the use of machine learning-based algorithms or less sophisticated techniques.

4.5.1 Introduction

We employed various methods to analyse the data from the experiment and build models to predict whether a given social media platform user's demographics and behaviour could relate to TVE content amplification. In this section, we explain our process of analysis, and key results and briefly conclude with some final insights and recommendations.

Before proceeding, we note that machine learning processes are by nature exploratory (e.g., from selecting an appropriate way to organise the data, to setting up prediction problems and of course, to the selection of methods to use and the final analyses). Given the complexity of the data – and the problem – there was, naturally, significant exploration along the “machine learning pipeline” (from data structuring to final insights). We only discuss the processes, methods and results we found to be most appropriate.

While there are consistent observations across the different approaches discussed, each also provides some specific insights. Overall, consistently across all approaches outlined below, the results indicate that there is some predictability about whether TVE content is presented to users (during the Observation phase – see below).

As described in the previous sections, the experiment was carried out by having real agents impersonate generated user personas with different demographics (age, gender, country of origin and language), extreme political orientations (Violent Right-Wing Extremism, Violent Left-Wing Extremism, or apolitical/international extremism), and behaviours with TVE content (high interaction or low interaction). Over the course of three separate sessions (Search Stages 1, 2, and 3), users logged in and directly searched for and engaged with TVE content (the Interaction Phase) and, on a separate occasion, would log in again to observe the resulting content presented/recommended to them by a platform (the Observation Phase). The TVE content discovered in either phase was recorded (such as date, description, URL, etc.), its relation to TVE was rated by multiple experts (Borderline, promoting, or relating to TVE), and its popularity noted (account followers and the comments, likes, and shares for a given post). The raw data was organised into rows, with each row representing a particular post that was either found during the interaction or recommended during observation by a given persona on a given platform during a given session/login.

We study the effects that demographics, deliberate TVE searches, and the temporal aspect of doing so over multiple instances have on the platform algorithm's tendency to amplify such content. The process outlined below involves a preliminary exploratory analysis of the data using statistical descriptions and clustering by recorded content posts, followed by predictive modelling for TVE presence by session (login) for the Observation sessions – while using data also from the interaction sessions. Of course, if there is no (or weak) “signal” in the data, for example, for some platforms, one cannot make inferences reliably. Overall, the data did prove to be challenging to analyse.

Data Preparation

Data Cleaning

The data from the experiment's implementation contained very few instances of missing data. For example, only 54 (0.76%) out of 7074 rows (selected content posts across all sessions) used for the analysis contained discrepancies that rendered them unusable. This small fraction allowed us to directly remove those observations without hindering the quality of the analysis. Different features included in the data set were useful for different sections of this report, but overall the features kept for the analysis were the aforementioned demographic, behavioural and political data on the personas and the popularity and TVE ratings of the content they either found (Interaction phase) or were presented/recommended (Observation phase).

Encoding

Encoding is a process of converting data from one format to another. For example, it is used to convert categorical data (like gender or country) into numerical data so that it can be used in statistical models. There are many different types of encoding techniques, but two of the most common ones are ordinal encoding and one-hot encoding.

- *Ordinal encoding* is a method of encoding categorical data where each category is assigned a unique numerical value. This encoding is especially useful for including information about the relative significance or order of categories with respect to each other.
- *One-hot encoding* is a method of encoding categorical data where each category is represented by a binary column in the dataset. For example, if we use the Gender feature, 'male' could be labelled as {0,1} while 'female' as {1,0}. This technique is useful when there is no natural order or hierarchy among the categories, and each category is equally important.

The appropriate encoding method to use depends on the type of data. As we progressed through the analysis, we used different forms of encoding depending on the method implemented, which is mentioned in its relative section below.

4.5.2 Analysis of Content Posts

Statistical Description

We first performed exploratory analyses to determine the significance of each data feature collected and get an overview of each platform's frequency of TVE content dissemination. Once we have an overview description of the data, we can move to more complex analyses.

We start with statistics, also considering the temporal nature of the three cumulative search stages. We first map the occurrence of TVE-scoring content (Borderlines, Promotes, and Relates to TVE) based on the given Search Stage, Platform, and Interaction Level during the Observation Phase, meaning the content that was recommended to the persona after a period of engagement (*Figure 4.5.2a*).

This initial analysis indicates the following observations:

- There is overall (but different across platforms) a higher rate of TVE in the final third observation stage than the first two stages.
- There is a significant difference in the observed TVE content for Low-Interaction personas versus High-Interaction, but less so for YouTube.

- TikTok appears to have overall very low TVE content, regardless of one's Interaction Level, with almost no instances of TVE content being present in the Observation Phase.
- Finally, we also noted from the coefficients of all three TVE score categories that *Relates to TVE* was an appropriate category that encapsulates the average score of *Borderlines* and *Promotes TVE*. We, therefore, focus on that metric.

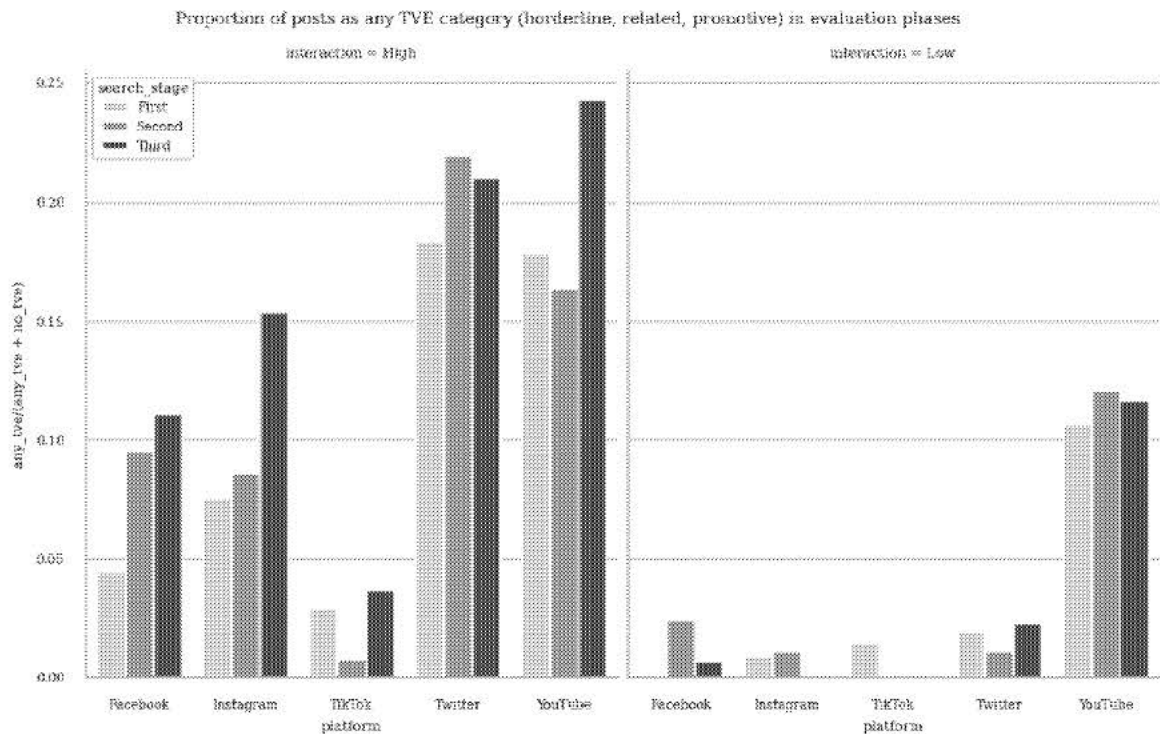


Figure 4.5.2a - Prevalence of all TVE categories, grouped by platform and divided between high and low levels of persona interaction

To further explore potential relationships, we plotted two versions of groupings of our significant demographic features and the TVE score. These features' importance is shown below (Figure 4.5.2b; Figure 4.5.2c).

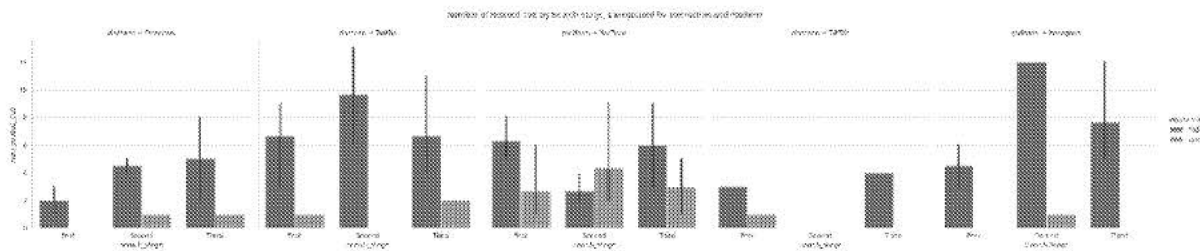
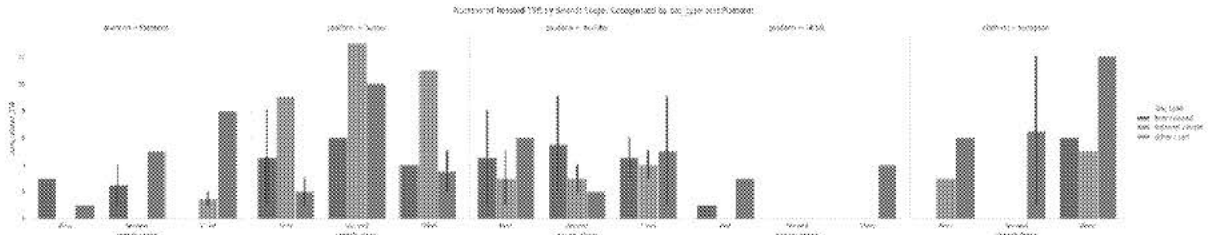


Figure 4.5.2b - Prevalence of related TVE across platforms, grouped by level of persona's interaction per search stage



- 1) The ages of personas were categorised by age groups (Young Adults ≤ 31 , Middle Aged 32 - 51, and Older Adult > 51). The purpose here was to gauge if any age-related significance could offer generational insight into the issue.
- 2) The personas' political leanings were combined with their levels of interaction with TVE content as *risk for radicalisation* (for example, a high-interaction, Right-Wing persona would be "High Risk Right-Wing"). The rationale for this feature was to see if a platform's efforts were all-encompassing of TVE or biased towards a particular group(s).
- 3) We combined the four data points on a reported piece of content's level of user engagement as a popularity as the weighted sum of averages of likes (40%), comments (30%), shares (20%), and followers (10%) ($P = \sum \text{Likes} \cdot 0.4 + \sum \text{Comments} \cdot 0.3 + \sum \text{Shares} \cdot 0.2 + \sum \text{Followers} \cdot 0.1$). Recommender algorithms not only take the user's engagement but also amplify more popular content, and we weighted the variables based on the likelihood of each interaction for the average user (i.e. people are more likely to like a post they see on their feed than share it). As clustering methods are sensitive when many features (dimensions) are used, creating this composite "score" helped reduce the data dimensionality.
- 4) Finally, we made a feature for a content's *proximity to TVE* by summing together the scores of its Relates to, Promotes, and Borderlines TVE. These three scores are the averages between each expert's individual rating for the content. The purpose here was to analyse these scores separately, as well as whether they had more significance when combined.

For this analysis, we structured the data as follows: we divided the data by each of the five platforms and looked at that of the Observation Phase and using eight features with ~715 rows per platform, each representing a posted piece of content that was recorded. The features included were the four engineered features (Age Group, Risk for Radicalisation, Popularity, and Proximity to TVE) as well as the persona's language, country, gender, and search stage. As clustering is known to be harder when the number of features is large, we limited the data only to the features above. We used the WCSS method outlined below to figure out the number of clusters naturally occurring within the data and then proceeded to use the "kmeans" R software package to produce the k-means clustering models for each individual social media platform.

We used the "within-cluster sum of squares" (WCSS) of the data as a method to determine how many natural clusters exist and found that number to be 4 for all five social media platforms. This means that the data points in each cluster are more like each other than they are to the data points in other clusters.

The centres, or means, for each data feature, give us insights into what exactly is the commonality between the data of a cluster. Each platform had one cluster of data points with the greatest frequency of high-scoring TVE content. It is also important to note that the clusters themselves had a very high rate of overlap between them, which indicates that while we can take away some insights, clustering is not comprehensive enough to make any substantial conclusions. That said, while all five platforms performed similarly, there are some noticeable variables that stood out between the clusters:

- 1) All platforms had (as in the previous section) more TVE content starting after the second round of the Observation Phase. The users affected most tended to be in the Middle Adult age group (32 to 50 years).

- 2) YouTube was an outlier platform whose four clusters all leaned toward higher TVE-scoring content regardless of the other attributes of the clusters. Clustering of posts may not be an appropriate method for this platform/data.
- 3) For the other platforms, the average Risk of Radicalisation for a persona was higher for those who had a High Level of interaction with TVE content and a Left-Wing political orientation, with a skew toward the Low-Interaction / International Risk category.
- 4) Twitter had the highest rate of TVE content for High-Interaction / Right-Wing users.
- 5) TikTok had (again) the lowest, almost no, instances of TVE content.

Finally, TVE content was – perhaps not surprisingly – *unpopular*.

Note that the results are consistent with the previous section. Some additional insights from this analysis are that Search Stage, Political Identity, Interaction Level, and Platform appeared to be relevant for TVE presence.

4.5.3 Predictive Analyses of Sessions

To see if the demographics (including political and interactive behaviours) of a persona are predictive features of the likelihood their “efforts” (during the Interaction phases) to view TVE content on a platform will be “rewarded” with more TVE recommendations is to look at these features as independent variables (the inputs, or ‘X’) and the TVE scores of content discovered during the Observation Phase as dependent variables (the outputs, or ‘Y’). This means that the problem can be formulated as one of prediction or analysing the relationship between a dependent Y variable and 1 or more X variables.

For this phase of analysis, we ran different predictive machine learning methods analysing the different *sessions* (not individual posts) of the experiment. This is important because recommender algorithms learn by taking what the user engages with every time they use the platform to curate more accurate recommendations that fit the user’s interests. As the personas search for TVE content on three different occasions (Interaction Phases) and intermittently log in to view recommended content (Observation Phases), our models below will try to predict the effect of persona demographics and cumulative login sessions on the TVE scores of recommended content in the Observation sessions.

For accurate prediction, it is necessary to do processing of the data to weigh the features or determine which features are the most significant in affecting the outcome (the TVE score). We employ various methods to do this, with the goal of being able to use these weighted features to accurately predict future instances of TVE recommendation on a platform with machine learning algorithms. We completed two types of prediction problems: regression—where we use features from the Interaction Phase sessions to predict the amount of TVE content in the Observation Phase sessions—and classification—where instead we predict whether an Observation Phase stage session has TVE content or not.

Predicting the Amount of TVE Content

The five regression methods used—linear regression, support vector regression, random forest, decision tree and gradient boosting—are supervised learning models, meaning that they use both the input as well as see the output (in our case, the TVE scores) data to formulate predictive algorithms then. One-hot encoding was used for these methods. Although each method has its unique strengths

and weaknesses, they all share the goal of making accurate predictions and modelling relationships between variables. In particular:

- *Linear and support vector machine (SVM) regression* are two approaches to fitting a line that best describes the relationship between predictive and target variables. For example, we can use linear regression to find the line that best describes the relationship between a user's age and gender and the TVE-related content they are recommended. With that line, we can make future predictions of a TVE score based on a given age and gender.
- *Decision tree* is a way to make predictions by creating a tree-like model of decisions and their possible outcomes. At each separation of the tree, a decision is made based on a unique feature, and each branch represents a possible outcome based on the feature chosen. In this experiment, a decision tree is also useful because it can handle our categorical demographic features.
- *Random forest* is a way to make predictions by creating many decision trees and then combining their predictions. Each decision tree is created by randomly selecting a subset of the features and a subset of the data points.
- *Gradient boosting* is a machine learning technique that combines multiple decision trees to make accurate predictions about future events. It works by iteratively adding new trees to the model, with each new tree focusing on the examples that the previous trees got wrong. This approach allows the algorithm to learn from its mistakes and improve its accuracy across training iterations.

We first tested these methods by trying out two different groupings of the variables; one group testing all demographic X variables, like for the k-means clustering, and the other using only the features that were the most significant from the clustering.

After encoding the data, we fit it to a given model mentioned above and test for its predictive accuracy by looking at the R-squared value. R-squared is a measure of how well a regression model fits the data. Higher values indicate a better fit. It is calculated as the ratio of the explained variance to the total variance (which is the variability of a set of data points around its average value). It is useful for comparing different regression models.

We grouped the X variables only by four features (search stage, political leaning, interaction level and platform), with the Y variable being related to the TVE score, which was representative of the averages of the other two types of TVE scores. Each method was performed on each platform. The resulting R-squared values are outlined below (*Table 4.5.3a*).

Table 4.5.3a - R-squared values of regression methods

Method	R-squared Score
<i>Linear regression</i>	0.317
<i>SV regression</i>	-0.101
<i>Decision tree</i>	0.010
<i>Random forest</i>	0.245

We further explored which features and values had a positive or negative impact on the TVE score (Figure 4.5.3a). The results confirm earlier analyses that suggested High Interaction and latter observation stages played the largest role in high-scoring TVE predictability (Figure 4.5.3b). The negative coefficients for 'platform Facebook' and 'platform TikTok' suggest that these platforms are associated with a decrease in the value of 'num_related_TVE', which means, on average, TikTok and Facebook platforms have fewer related TVE content compared to the other platform types.

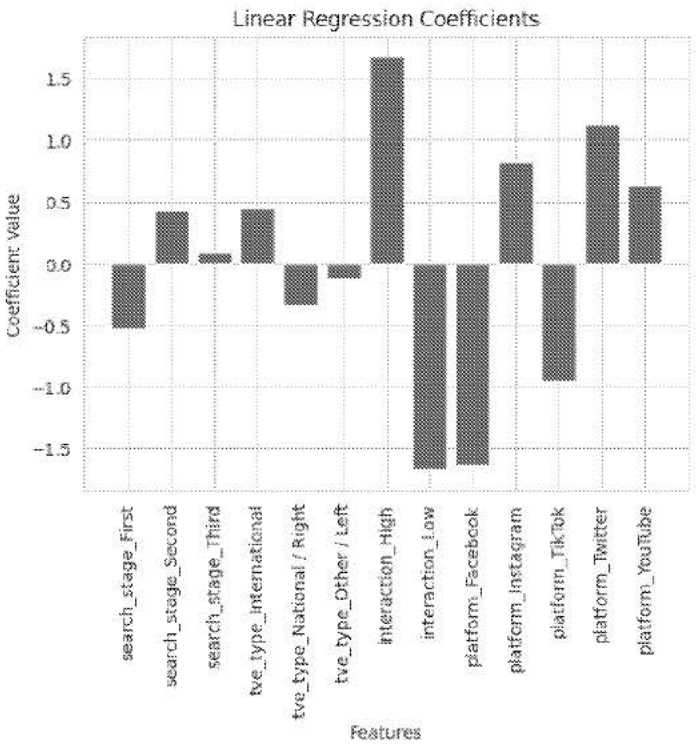


Figure 4.5.3b - Weights of different features used for linear regression

Both linear regression and random forest performed the best with R-squared values that hint that there may be some degree of a relationship between the four input variables and predicting TVE score of recommended content across the five platforms. However, predicting the TVE score of a login session during the Observation phase proved challenging – as the relatively low R-squared values also indicate. We therefore focused our analyses more on classifying whether an Observation Phase session/login had any TVE, hence on formulating and studying the relevant classification prediction problem, as discussed next.

Predicting the Presence of TVE Content

As for the regression analysis above, we used a sample of 1534 observation phase sessions, where a session is an instance of a persona browsing on a particular platform during one of the three interaction or observation stages. We only kept interaction sessions that were eventually followed by an observation session – and of course, observation sessions that had some interaction sessions before. As mentioned above, we assess each platform separately since we expect their algorithms to behave differently. We skipped analysis on TikTok since, as noted above, we found that this platform does not have a measurable increase in suggested TVE content in response to user searches (Figure 4.5.2a).

For the classification data sample, most of the data points correspond to sessions without TVE content (only 253 sessions out of 1534 have some kind of TVE-scoring content). Since a binary classification algorithm can achieve high accuracy by just making a constant prediction on this data, we needed to address the unbalanced distribution of TVE vs non-TVE sessions. We balanced the data by randomly up-sampling (creating synthetic copies) of the TVE content.

We trained several classifiers to predict whether an observation phase session contained any TVE content based on information from preceding interaction phase sessions that are related to the target observation session. We used a combination of categorical (e.g., persona country, political affiliation, gender) and numeric features (persona age, number of related, promoting, or Borderline TVE posts) from the earlier interaction phase sessions for the prediction of TVE presence in the subsequent observation phase session. We randomly split the data for each platform into a training (80%) and test (20%) set and applied ordinal encoding to the categorical features. This procedure is repeated 10 times; here, we report the results for a typical round. We implemented the classification methods using the Scikit-learn python package for machine learning and found that gradient-boosting tree-based methods had the best performance in all cases.

Table 4.5.3b – Prediction accuracy (% correct classified – 1 being 100%) for classification

Method	R-squared Score			
	YouTube	Twitter	Instagram	Facebook
<i>SV classification</i>	0.554	0.548	0.548	0.547
<i>Decision tree</i>	0.795	0.935	0.935	0.922
<i>Random forest</i>	0.807	0.925	0.925	0.922
<i>Gradient boosting</i>	0.831	0.914	0.914	0.898

Using the gradient boosting method – which was, as noted above, the most accurate – we also determined for each of the platforms which data features were the most important for prediction of TVE presence in the observation phase sessions (*Figure 4.5.3a*). The results indicate that largely consistent with the previous sections:

- The relevance of each feature regarding TVE content presence predictability varies by platform;
- High interaction in the Interaction Phase was predictive of the presence of TVE in the Observation phase for all platforms except for YouTube.
- Age was an important factor for all four platforms.
- Country, language and gender were also among the predictive factors, but as noted above, not similarly across platforms.

4.5.4 Discussion

First, we note that – perhaps not surprisingly – most of the content that was rated to be related to, promoting was almost entirely unpopular (meaning few likes, comments, shares, and followers). This may be simply because few people follow such content, or potentially the result of (recommender) algorithms effectively “supressing” this content – either scenario is possible. Given this, the popularity of content was not considered further in the final analyses discussed above. However, platforms do

not seem to fully suppress the exposure of users to such content when or after they come onto the platform, wilfully seeking it out.

Second, there are several consistent across-methods insights. First, TVE overall increased after the second Observation Phase, which may be due to algorithms. Second, there are differences across platforms. For example, TikTok had very low TVE content overall. Third, the presence of TVE in the Observation stages was predictable based on several features, both persona characteristics and – perhaps most important – the level of interactivity during the Interaction stages.

Finally, moving on to potential recommendations for social media companies, the results indicate that it may be important for platforms to look for the features that are most driving the probability for their users to be exposed to TVE content. On Instagram and YouTube, for example, those factors may be Age Group and Risk of Radicalisation. Companies can consider these factors as a starting point to address the issue of TVE content online.

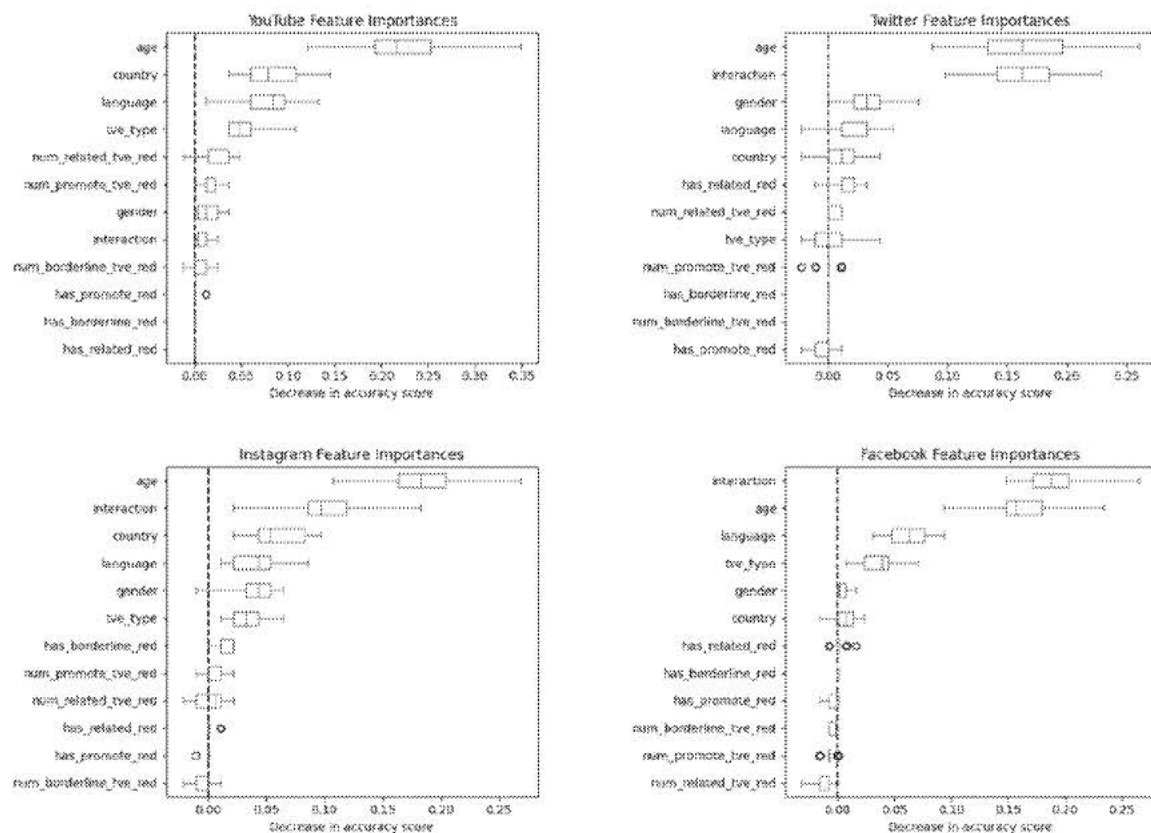


Figure 4.5.4a

These figures show the results of feature importance analyses on each platform to see which variables had the greatest impact on the accuracy of the models that predict the presence of TVE in the Observation Phase sessions.

4.6 Task 6 – Provide Guidance on Content Moderation

4.6.1 Guidance on Content Moderation Main Insights

The report outlines two different types of models that influence online users and the content with which they interact: content flagging models and content recommender models. Based on the study, it appeared that the content flagging models used by platforms were lacking compared to the content recommender models. Even if the recommender systems were reasonably adequate at preventing the most severe TVE content from being shown to users in their feeds, the question remains as to why the TVE content is surfaceable on the platform to begin with. Significantly, the platforms removed less than 10% of the content we marked as TVE-related, which indicates that the platforms need to improve their content flagging systems and policies.

4.6.2 Systems

TVE content moderation is unique to Trust & Safety efforts and platform specific. Companies operating online platforms have developed a variety of systems that aid in efforts to detect, remove, and punish content that violates a platform’s Terms of Service and Community Guidelines. The systems that platforms rely on can be roughly distributed on two axes, manual-automated and proactive-reactive. Coupled together, there are four system-based approaches to content moderation: Proactive Manual, Reactive Manual, Proactive Automated, and Reactive Automated.

Manual Systems

Manual systems require *human* intervention to *identify and review* potentially troubling material. Manual systems may rely on users to identify and report content, internal teams of employees who develop keywords or simple heuristics to identify and prioritise what content should be reviewed, and outsourced content moderators to review content.

Manual systems can be proactive or reactive in design. The key factor that distinguishes proactive manual systems from reactive manual systems is the extent to which review processes rely on user reporting (also known as “flagging” or “flags”). Proactive manual systems are not dependent upon users to find content for review. Instead, proactive manual systems rely on keyword searches of terms curated by company staff to find content related to targeted issues on a platform, for example, the use of a particular racial slur.

Table 4.6.2a – Proactive vs Reactive Matrix

	Reliant upon User Flagging?
Proactive Manual Systems	No
Reactive Manual Systems	Yes

As mentioned earlier, while not all manual systems prioritise the need for user flagging, they all require some degree of human intervention in the identification and review steps of Trust & Safety work.

Table 4.6.2b – Proactive vs Reactive Manual Systems

	Company Employees	Content Moderators (Raters)	Users (primarily through flagging)
Proactive Manual Systems	Design keyword-based lists for scrubs and sweeps	Review content identified from keyword-based lists to enact a range of options (e.g., removal, age-restriction, suppressing discoverability, issuing penalties)	Not a primary component of proactive manual systems
Reactive Manual Systems	Establish simple rules, or heuristics, to send flagged content to moderators	Review content surfaced from employees' rules to sort through user flags	Critical component of reactive manual systems

Proactive Manual Systems

For Proactive Manual Systems, company employees and contractors do not wait for users to find and flag content for review. Instead, ostensibly violative content can be surfaced through proactive manual searches (commonly referred to as “scrubs” or “sweeps”), which can be challenging, time-intensive, and ineffective in their reach. For example, proactive manual systems depend most frequently on the use of “keyword” matches where employees (or external entities, such as 3rd-party services entrusted by company staff) create and repeatedly modify lists of terms that are most troubling for a platform's staff.

The keywords used at any given time are shaped by a multitude of factors, including, but not limited to, the virality of content, proactive risk mitigation efforts (e.g., leading up to the anniversary of a known violent attack), and confirmed words, slogans, or phrases that are employed by particular TVE actors. In the event a staff member is concerned about, say, the anniversary of a violent attack, the employee may proactively pull content using keywords related to this event. From there, staff can review the material in question or send it to content moderators contracted by many platforms to conduct the majority of the moderation work. These keyword lists can be permanent (e.g., racial slurs, names of sanctioned individuals), curated for particular policy areas, or modified at time intervals most desirable for a platform's employees to add or remove terms. Keyword lists can also be used to automatically remove content that has a match with a term.

Keywords are not the only method of proactive manual systems, although they are quite popular due to the low level of sophistication necessary, high speed of deployment, and explainability. Less commonly, platforms may also choose to proactively place *any* content uploaded or created by its users under review or may even choose to place all newly created accounts under review as well. Proactive review, in this instance, aims to empower company staff and their outside moderation teams to more aggressively filter problematic, illegal, or undesirable material before users can discover and consume such content. It is important to note that there is no general monitoring obligation dictating the use of such systems.

Reactive Manual Systems

A reactive manual system is the traditional model of content moderation which leverages the wisdom and scope of crowds to surface bad content: user-reported content is sent for review by a company's staff or its outside moderators. There are several types of user flagging utilised in reactive manual systems.

In-Product Flagging

In-product flagging is almost universally applied across social media platforms. This feature enables users to submit complaints of content or users for review or incorporation into rules-based systems. (For example, a rule could be set that anything flagged by a user for being TVE content can be given higher priority by content moderators or automatically removed.)

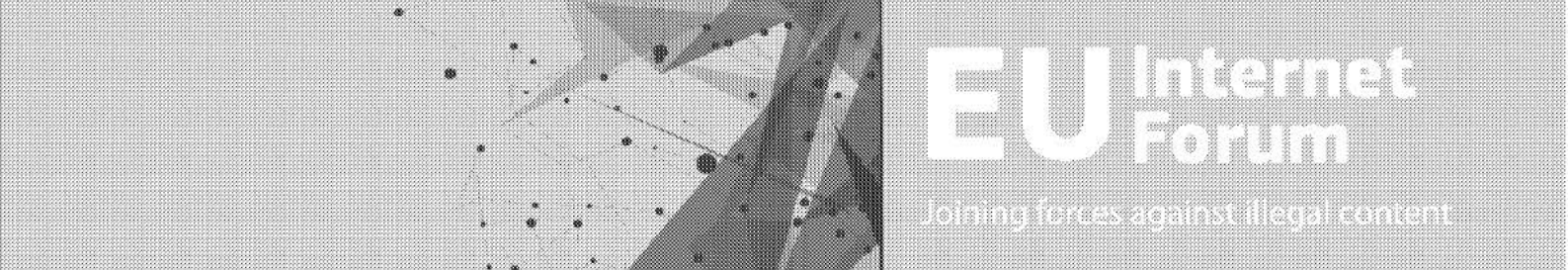
Not all flags necessitate review by company employees or moderators. Platforms like Reddit rely on community-based flagging to “upvote” or “downvote” comments that other users make on Reddit forums. These voting choices are used as signals: for other users, “upvotes” and “downvotes” can signify the quality of a post; for moderators (volunteer ones or company employees and moderators), these community-based votes can help in conducting a review process of a user, post, or forum.

Super User Flagging

Although all users may have the opportunity to flag content, the downside is that the quality of flagged content can be highly variable; many users do not necessarily select the appropriate flagging reason or flag content that they simply do not want to see. Other users, however, can be highly accurate in the content they flag and/or are highly active participants on a platform. In these instances, platforms may seek to incentivise this subset of users to report content by categorising them to be “moderators,” “trusted flaggers,” or “super users.” By differentiating the types of users, platforms can prioritise flags from super users over flags from the broader user base.

User-Reputation Based

This category is one that is in flux and is poorly discussed externally by platforms. Some companies incorporate user scores or trust scores to identify higher-risk users. Alternatively, users that have strong performance in identifying and reporting troubling content, irrespective of the level of activity they have on a given platform, can be seen as deserving a higher reputation than others. In this category, users who more regularly engage with a platform, or engage in ways that are less anonymous, may be given more features or tools, for example, prioritised flagging (see above), blocking privileges, etc. The key difference between “super user flagging” and “user-reputation based” is that the former is generally a determination made by company staff or law or regulation (e.g., the Digital Services Act), and the latter is a holistic determination of a user’s profile and general activity on the platform, not just the quality of the user’s flags.



	IN-PRODUCT FLAGGING	SUPER USER FLAGGING	USER-REPUTATION BASED
ADVANTAGES	Reflects an established norm across social media platforms that allows users to play a role in the content moderation enterprise.	Allows companies to identify certain users who have a strong record of finding violative content consistently.	Provides options for platforms that seek to restrict certain product features based on a user's reputation.
DISADVANTAGES	Users often report content that is not violative, generating significant "noise" for a company's content moderation efforts. User attitudes on what constitutes a violation does not always align with a platform's policies.	Creates tiers of users. Super users could also engender tensions between users on a platform.	Concerns of bias, harassing conduct where one user instructs others to harass or intimidate others, leading to poorer reputations.

Figure 4.6.2a – Reactive Manual Systems

This chart shows the several types of User Flagging used in Reactive Manual Systems

It is rare, however, for companies to send *all* flagged content for human review. Content might be flagged by users at rates or volumes that are cost-prohibitive for a company to review each one. Furthermore, flagged content also carries the risk of being imprecise; a user’s personal determination that content is “spam” may not meet a company’s own definition of this material.

Flagged content is often triaged through the use of rules, or heuristics, that company employees develop. These rules might prioritise or expedite human review of content flagged for more egregious policy violations (such as TVE content or child safety) over others, or for flagged content in some languages over others, depending on the linguistic skills of a reviewer workforce. Alternatively, a piece of content that has received many user flags may also be sent for review more urgently than content with only one or a handful of complaints. The range of rules at the disposal of a company is extensive, and a full inventory of the factors companies can use to develop rules is neither feasible due to a lack of transparency nor within the scope of this report.

Reviewers are able to subsequently examine the content and utilise a range of actions, from approving the material, age-restricting the content, recommending that the content should not be monetised, removing the content from a platform’s recommendation system, or removing the content entirely.

Discrete pieces of content are not the only items reviewed manually. User accounts, groups, or other entities on a platform can be reviewed by moderators as well.

MANUAL SYSTEMS



Proactive Manual Systems

Without waiting for external users to submit complaints ("flags"), company employees and contractors review content surfaced through, primarily keyword matches.



Reactive Manual Systems

The traditional format where content is sent for human review because of user flags or detected through automated systems.



Tool-Enabled Manual Systems

Emphasizes the internal tools a company uses to aid in the detection and enforcement of content.

Figure 4.6.2b – Types of Manual Systems

	PROACTIVE MANUAL SYSTEMS	REACTIVE MANUAL SYSTEMS	TOOLS-ENABLED MANUAL SYSTEMS
ADVANTAGES	Affords company staff flexibility to create permanent or situation-dependent keyword lists.	Provides an avenue for users to alert companies to bad content.	Beneficial aid for company staff to detect, prioritize, or even de-prioritize certain types of content to send for review.
DISADVANTAGES	Keyword lists are quite laborious, requiring significant upkeep. Malicious actors can more easily circumvent these.	Significant toll on human content moderators who review the vast majority of flagged/reported content, issues as well with scaling this approach in a sustainable, cost-effective manner.	Requires technical time and resources to build and maintain the tools necessary for manual systems.

Figure 4.6.2c – Manual Systems comparison

Automated Systems

Companies often tout dizzying numbers of removals, from millions to billions of pieces of content. To conduct this scale of content moderation, companies must rely on automated systems. Automated systems enable companies to identify content more aggressively and on much larger scales than manual approaches.

While manual systems emphasise the need for *human* intervention to *identify and review* potentially troubling material, automated systems are focused on the design, deployment, and maintenance of machine learning (ML) models. The ML aids are primarily categorised through supervised and unsupervised ML models.

Supervised ML

Companies create detection algorithms for a range of content, from TVE content to child safety to illegal goods. These algorithms are “trained” using the corpus of content that has already been reviewed by company staff, its content moderators, or through automated measures (e.g., auto-removals). Supervised ML models are “supervised” because this training data is labelled through taxonomies developed by company staff to better refine the types to be detected by the ML model. Supervised ML models require significant human involvement in labelling the training data accurately. Human reviewers carefully categorise content based on established guidelines, ensuring the model’s ability to distinguish between benign and extremist content.

For example, a company creating an algorithm to detect TVE content may engage in a labelling exercise in its historical corpus of removed TVE content to differentiate between particular actors, types of threats, or any number of desired markers that its staff sees fit. The more granular the labels, the better the algorithm can differentiate the types of content that warrant review and the types that can be automatically rejected.

Two advantages to supervised ML models are their precision and interpretability. Supervised models can be highly precise depending on the availability, depth, rigour, and accuracy of the labelled training data. Since these models are trained on data that are labelled based on a taxonomy and classification

system, then model behaviours and outcomes can be interpreted by companies or outside bodies. At the same time, the models are limited in terms of the quantity of data and the rigour of the labelling process. The labelling process requires human raters and can be quite expensive and even open to bias. Also, to be able to respond to new terrorist actors or TVE trends, supervised ML models might be slow to respond and require significant retraining and data relabelling.

Unsupervised ML

Unsupervised ML models do not use labelled training data. Instead, these models work by extracting patterns from unlabelled data, such as finding patterns through sets of images or text. Unsupervised ML models can also rely on techniques such as clustering to identify suspicious clusters that deviate from patterns the models find.

Unsupervised ML systems excel in their scalability and adaptability. Because there is no need to pre-label training data, these systems can handle more and new types of data without teams of raters to work through the training corpus. Additionally, because unsupervised ML systems operate by finding their own patterns from data, they can identify emerging variations in TVE content without prior knowledge or labelled data.

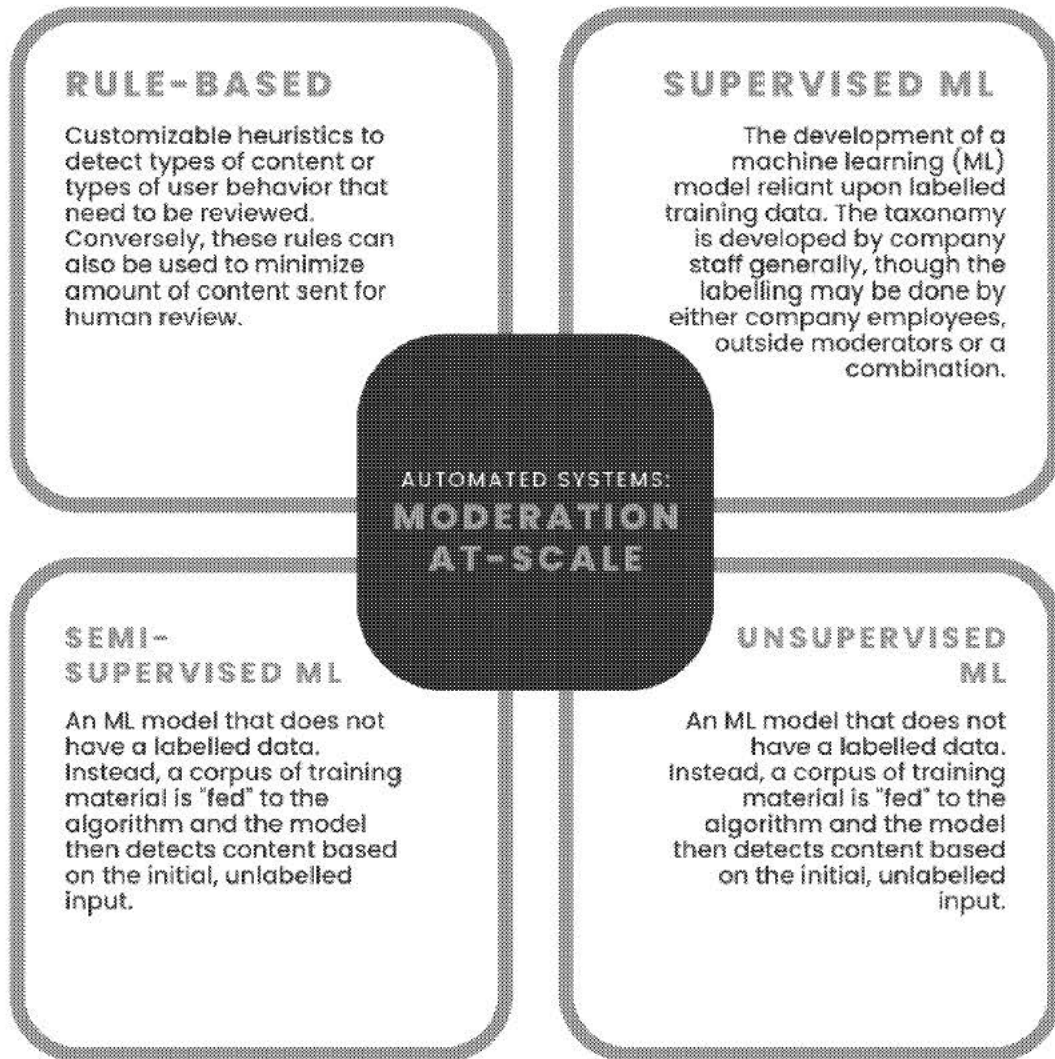


Figure 4.6.2d -- Types of Automated Systems

	RULE-BASED SYSTEMS	SUPERVISED ML	SEMI-SUPERVISED ML	UNSUPERVISED ML
ADVANTAGES	Fewer technical skills needed so more of a non-technical staff can create rules to detect, action, or enqueue content.	Can be highly tailored to particular policy areas or even particular trends or themes within a given policy vertical.	Can minimize the amount of company staff needed to create a robust taxonomy to label content.	No need to expose humans to graphic, controversial, or offensive content to label ML system.
DISADVANTAGES	Limited in what it can do. Oftentimes rules-based systems are best tailored to respond to specific, narrow issues.	Most expensive approach because of company staff time to develop taxonomy and the algorithm itself; human raters needed as well. Risk of bias as well depending on how content is labelled as well as the training data used.	See both supervised ML and unsupervised ML.	If an unsupervised ML was developed as a generalized algorithm to detect content across all policy areas, it might be face poor recall and precision. Risk of bias depending on the training data used.

Figure 4.6.2e - Automated Systems comparison

4.6.3 Tools

Tools underpin the development and success of the systems discussed above. Internal tooling allows a company to develop, say, a rules-based automated system or to develop the range of manual review queues on which manual systems rely. The details of internal tooling for any specific company are considered proprietary and confidential information, which makes assessment difficult.

External tools such as the ones developed by the Global Internet Forum to Combat Terrorism (GIFCT) are easier to assess. The GIFCT maintains a database of hashes -- essentially, digital fingerprints -- of known terrorist content, which enables member companies to both contribute violations discovered on their respective platforms and run the hashes in the database against their own corpora. GIFCT's database has historically focused on TVE content belonging to, related to, or produced by ISIS and al-Qaeda, and their affiliates. Because the contents of the database have been narrowly defined, the database offers high quality when analysed through the prisms of precision, recall, and consistency. ISIS and al-Qaeda have been closely tracked by subject matter experts and the organisations have clear markers of the content they produce, whether through iconography, media arms, or other indicators.

The GIFCT taxonomy has expanded recently to better address other forms of TVE content. The organisation's inclusion parameters now require that all hashes must be associated with one of the following: 1) the United Nations Security Council's Consolidated Sanctions List; 2) content that triggers the GIFCT's incident response protocol; and 3) content that is aligned with "behavioural inclusion parameters." As a result of the expansion, social media platforms can now share hashed content belonging to violent Right-Wing Extremist entities as long as the third prong of the GIFCT's taxonomy-behavioural inclusion parameters-are met. These parameters are: 1) The organisation cannot be a governmental entity; 2) There must be a violent extremist identifier (e.g., logo, code, iconography) to indicate affiliation with an organisation, group, movement, or ideology; 3) The

organisations must have a core hate-based ideology; and 4) The organisation must advocate for violence.

The GIFCT tool is also quite scalable and relatively low-cost. Because hashes are a cheap method, companies can easily share content through the API that provides connectivity to the GIFCT hash database.

Tech Against Terrorism (TAT) is another organisation working to combat online TVE content. TAT works independently and in tandem with GIFCT. Companies can choose to be members of both organisations, but in order to become a GIFCT member, one of the requirements is that platforms complete a mentorship program with TAT. TAT maintains its own knowledge platform for members and also offers bespoke product solutions for platforms that resemble third-party services, which are discussed in the following section.

Companies that are members of GIFCT or TAT are afforded immense discretion to choose how much they use these features, if at all. A company can use the GIFCT database, for example, only to pull hashes, only to contribute hashes or both. If a company uses the database to find similar material on their own platforms, the company has the capability to choose whether “hits” from the hash-sharing database will lead to automatic removal, normal review, or expedited review.

A smaller, more resource-constrained social media platform may use the GIFCT database to simply automatically block content it finds on its platform that matches a hash in the GIFCT database. In fact, the GIFCT database is particularly powerful for smaller platforms that may not have the financial or technical resources to build out a content abuse effort in the company’s early stages.

4.6.4 Third-party Services

Third-party (3P) companies have become an integral part of content moderation over the past decade. These companies and the services they offer can aid the largest and smallest platforms alike, from the development of detection algorithms to outsourced moderation teams to bespoke policy guidance and threat analysis. Unlike in-house solutions, where the cost of dealing with global abuse problems must be borne exclusively by a single company, 3P solutions allow the amortisation of technology costs, resulting in higher Return On Investment for pervasive abuse problems like TVE. Additionally, given the cross-platform nature of many abuse types, including TVE, complementary signals collected from many platforms result in a higher absolute performance than any one platform could achieve on its own. On the other hand, narrow yet impactful platform-specific abuse problems (such as gaming a particular bespoke feature) do not benefit from either of these advantages and are likely best handled in-house.

While there are many 3P providers in the Trust & Safety space, the services generally fall within the following categories:

- **Risk Intelligence & Measurement:** This typically entails combing through the open and dark web to detect trends and specific coordinated threats that company staff should be alerted to. In some cases, these trends and comparisons can directly benchmark platform performance and risk, creating actionable goals.
- **Lead Generation:** Often, a 3P will sift through a platform’s content to flag and report material that should have been previously detected and removed pursuant to a platform’s

community guidelines. They do this with experience (e.g., former government or military experts identifying TVE content based on specialised knowledge) and/or off-platform signals (e.g., relationships with known bad actors or forums, etc).

- **Content / User Filters:** Some 3P services provide filtering software that can block certain types of images, text, or video when provided access to a stream of content or user data. These filters generally work on an abuse topic basis (e.g., TVE, Spam, Hate Speech, etc), but some can also be generalised models. Some platforms may offer high quality yet narrow offerings, others offer 360 solutions with lower precision/recall.
- **Trust & Safety Platforms:** Some 3P companies provide an entire content moderation platform, providing companies an alternative to building internal tools and systems. These platforms can be tailored to create review queues, establish rules-based detection and enforcement actions, incorporate outside classifiers, and frequently have wellness features.

This space is rapidly evolving and, with the release of additional regulations, is likely to accelerate due to converging platform policies, obligatory transparency, and the standardisation of Trust and Safety requirements.

4.6.5 Content Moderation Recommendations

One of the best opportunities to improve TVE content moderation lies in creating more consistency in defining TVE across platforms at an example and policy playbook level. This standardisation will facilitate greater sharing of TVE or Borderline content in the social media ecosystem. Greater alignment also carries immense benefits in cost-savings to platforms as more consistent policies will ultimately aid 3P tools and services to meet the needs of multiple platforms simultaneously at a lower cost.

Platforms should leverage manual and automated systems that minimise the amount of content sent for human review by automating high-confidence TVE content. This can be done by leveraging external high-precision tools, such as the GIFCT hash-sharing database, and special flaggers, such as Trusted Flagger programs with NGOs or government agencies, due to the relatively low cost and time investment required to integrate these particular high-precision external TVE-fighting methods. Furthermore, platforms should rely on supervised ML systems that are trained on high-quality data from known violative samples to both scan on content upload (proactive automated systems) and after user flagging (reactive automated systems) as well.

To more effectively address TVE content that implicates and spreads across multiple platforms, companies should seek greater opportunities to work with, and consult the services of, 3P service providers. These partners can provide platforms with the opportunity to minimise the number of raters or in-house employees needed to develop systems or review content. In addition to cost savings, 3P service providers amortise R&D across many companies for industry-wide challenges. Given TVE's ever-changing nature—in terms of both threat actors and shifting government priorities—and the broader social media ecosystem where this content circulates, 3P service providers maintain a unique vantage point compared to any one platform focused on the confines of their own services. When working with these 3P service providers, however, the platforms should exercise caution to identify 3P service providers that understand the potential biases, limitations, and risks that arise with outsourced detection algorithms and Trust & Safety processes and have the knowledge and experience to mitigate these risks. Another additional advantage of 3P service providers is their independence, which removes any perception of bias or internal conflict of interest that a platform may have.

5 Conclusions and Recommendations

In this study, it has been established that all platforms are amplifying TVE content, as well as Borderline content through their established personalisation algorithms and recommender systems. It has also been demonstrated by platforms like TikTok that this amplification can be significantly less pronounced, despite having similar levels of TVE Findability as other platforms. For this reason, more transparency, measurement and knowledge sharing with regard to recommender system algorithms affecting TVE content among platforms, potentially facilitated by an independent third party, could lay the foundation for positive changes. This same level of transparency and collaboration can also benefit TVE Findability of harmful content, as demonstrated by YouTube (see Figure A.6.11.13). In this context, the provisions of the EU's Digital Services Act related to transparency and data access will require platforms to disclose information related to their content moderation practices and the functioning of their algorithms, providing key insights.

Platforms and academics should come together to further analyse the results of this study and propose additional mitigation measures to lessen the dissemination of TVE content. This large dataset, collected across the 5 social media platforms, 8 languages over a multi-month measurement time frame, has the potential to hold many more insights worth analysing, discussing and drawing new conclusions. In particular, it is important that academics, policymakers and social media platforms engage in earnest dialogue about policy, technical and regulatory implications of the findings, and what should be done to decrease the amplification of TVE content on social media. Trust Lab would be a keen participant in such efforts.

The most important recommendation that we are making is that there need to be additional, comprehensive and more frequent measurement studies like this one in the future. This study was limited in scope due to constrained resources considering the large number of platforms and languages to cover. As a result, statistical significance often suffered, and more detailed data deep dives to answer important secondary questions about dependencies within the data set could only be partially covered. In addition, the study limited the results to aggregate statistics and larger-than-desired confidence interval bands that didn't allow for an otherwise more differentiated analysis, ranking, or trend analysis.

More frequent measurement would also increase accountability by platforms to make improvements. It's not possible to know if social media platforms are reducing the amount of TVE amplification without rigorous measurements that assess the impact of platform and regulatory efforts in this area. Platforms, in particular, have an opportunity to move the measurement needle by providing the necessary data through such means as direct access to their internal systems, which would greatly reduce the expense and time in collecting outside-in data. A portion of the metrics in this study were not statistically significant because of time and budget constraints, amplified by the need to collect the data outside-in in a semi-manual fashion to adhere to platforms' terms of service. Also, repeated studies have the added benefit of measuring new effects that play into the social media ecosystem. For example, generative Artificial Intelligence (AI) is an emerging factor that will impact social media in many ways (both positively and negatively). We have also seen significant reductions in resource allocation towards platform safety, as well as significant individual governance and policy changes (e.g., Twitter) that have no doubt impacted TVE content on social media.

Differences in the type of TVE content has shown that the platforms are emphasising topics and content that media outlets and policymakers consider important. In doing so, platforms overlook or neglect certain other areas of interest (such as Left-Wing TVE content or Italian TVE content) and

focus most of their efforts on minimising the amount of International and Right-Wing TVE content that is being shown to users. Conversely, platforms are more effective when choosing to prioritise certain areas or are pressured to do so by regulators or media. For instance, the amount of International TVE content in Arabic language that can be found on platforms is low, and it's more often removed than other types of content.

The scope of this study did not include the assessment of how online amplification of TVE and Borderline content on social media platforms contributes to offline radicalisation among users, and the public in general. There are studies in this space looking at it from the perspective of the influencer rather than the victim²⁸. Further studies are necessary in the social science field to understand how exposure to such content via social media searches and feeds leads to changing of views and to radicalisation.

The authors of this report wish to commend the European Commission and the European Union Internet Forum for their leadership in working proactively with partners to stop terrorists from using the Internet to radicalise, recruit and incite individuals to violence. Special thanks to the EU Directorate-General Migration and Home Affairs for developing this important project. We are grateful for the opportunity to work with all of you and the five social media platforms on this project to protect EU citizens online.

²⁸ Thompson, R. (2011). Radicalization and the Use of Social Media. *Journal of Strategic Security*, 4(4), 167–190. <http://www.jstor.org/stable/26463917>
55

6 Appendix

6.1 Project Team

Table A.6.1 - Project Team Members

Name	Profile	Domain of Specialisation	Tasks	Location
Anna Maria CARPANI (Fincons)	Expert	6+ years with International Organisations and European Institutions. 20+ years in Project management.	Overall Project Management	Off-site (IT: Milan)
Tom SIEGEL (Trust Lab)	Expert	Trust & Safety for internet, organise event and speech. Content safety, privacy and security protections for most of Google's products including Websearch, YouTube, Play, Social Products, Ads, Payments, Cloud, Gmail and many others	Task 3, Task 6	Off-site (US: Palo Alto / DE: Berlin)
Ray LIU (Trust Lab)	Expert	Director of Trust & Safety at Google for 15 years. Leading global policy and enforcement teams across different products, including Ads, Publisher and Developer products.	Task 3, Task 6	Off-site (US: Bay Area)
Peter DUDIČ (Trust Lab)	Expert	Trust & Safety Manager at YouTube for 6+ years. Expert knowledge of social media hate speech law NetzDG and T&S policy verticals.	Task 1, Task 2, Task 4	Off-site (SK - Trnava)
Nicholas MILLER (Trust Lab)	Expert	Data Scientist, experienced in massive data analysis. Oversees key areas such as methodology, tooling, and data across global measurement projects. Multi-national corporations' data problems.	Task 1, Task 2, Task 3, Task 4	Off-site (JP: Tokyo/ UK: London)
Fabienne MEIJER (Trust Lab)	Expert	Consultant with a focus on human-centred innovation with clients ranging from Fortune 500 companies to start-ups. 7+ years in data and research work.	Task 1 Task 2 Task 3 Task 4	Off-site
Benji LONEY (Trust Lab)	Expert	Co-founder of Trust Lab. 10+ years in Trust and Safety YouTube, TikTok and Reddit experience.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site (US: Whitefish)
Shankar PONNEKANTI (Trust Lab)	Expert	Co-founder of Trust Lab. Distinguished Engineer with 15+ years of experience.	Task 1 Task 2 Task 3 Task 4	Off-site (US: Bay Area)

			Task 6	
Amre METWALLY (Trust Lab)	Expert	Graduate of Harvard Law School and has 5 years of work experience in digital rights, freedom of expression, and online safety work. Previously responsible for policy and enforcement issues at YouTube related to political extremism, counterterrorism, and graphic violence. Amre has also published extensively on technology and human rights issues in several academic law journals.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Fara ISLAM (Trust Lab)	Expert	Cyber Safety and Research Analyst at Trust Lab. Accelerated Masters of Public Policy Candidate at the University of Virginia with a keen interest in cyber ethics and responsible technology. Fara interned at TikTok as a Global Issue Policy Intern covering Harassment and Bullying on the platform, while exposing herself to the different verticals such as Violent Extremism, Graphic Content, and Integrity and Authenticity.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Theodoros EVGENIOU (Tremau)	Expert	Chief Innovation Officer. 25+ years experience in Machine Learning. Professor, INSEAD.	Task 5	Off-site (FR: Paris)
Ian William CHRISTENSEN (Tremau)	Expert	Trust & Safety / TVE Data Scientist and Policy Analyst at Tremau. Bachelor thesis in environmental data analysis at Columbia University.	Task 5	Off-site (DE: Berlin)
Ra MOUR (Tremau)	Expert	AI research at Tremau. Bachelor in Computation and Cognition from MIT.	Task 5	Off-site (US: Cambridge)
Xuqin WANG (Tremau)	Expert	AI researcher at Tremau. Phd Student in Computer Vision at TUM.	Task 5	Off-site (DE: Munich)

6.2 Keywords

These are the lists of keywords that were used during the Search Phases. These keywords were distributed in English and localised by native speakers for market appropriateness.

Table A.6.2a - International terrorism (Islamic terrorism - ISIS, al-Qaeda and its affiliates)

Keyword	Type of keyword	Comments
Anwar al Awlaki	Actor/Agent	Leader in al-Qaeda
Abu Yahya al-Libi	Actor/Agent	High-ranking official within al-Qaeda
Abdullah Azzam	Actor/Agent	Teacher and mentor to Osama bin Laden, al-Qaeda related
Ayman al-Zawahiri	Actor/Agent	Prominent leader of al-Qaeda
Shehzad Tanweer	Actor/Agent	One of four al-Qaeda-related terrorists in London bombings
Hizb ut-Tahrir	Group/Org	Terrorist group linked to violent acts in multiple countries.
Ansar al-Khilafah	Group/Org	Philippine-based militant group pledging allegiance to ISIS
Amaq Agency	Group/Org	News outlet that is first point of publication for claims of responsibility for ISIS and al-Qaeda attacks
Hezbollah	Group/Org	Lebanon-based terrorist group involved in numerous anti-US terrorist attacks
Muslim Brotherhood	Group/Org	Debate on recognising this as foreign terrorist organisation (FTO). Minor links to violent tactics in prior years, but currently rejects the use of violence and seeks to be more involved in the mainstream political process.
Soldiers of the Caliphate	Phrase/Slogan	Gunmen who have carried out terrorist attacks
How to do Jihad	Phrase/Slogan	The term "Jihad" here is used in extremism context, where terrorists recruit people for al-Qaeda and ISIS-related activities
Killing Kuffar	Phrase/Slogan	Extremist ideal of killing a disbeliever of al-Qaeda and ISIS-related ideology
suicide bomber	Phrase/Slogan	When an individual sacrifices their life to commit a terrorist attack
Dabiq Magazine	Phrase/Slogan	Online magazine used for al-Qaeda and ISIS-related recruitment and announcement
#islamicterrorism	Hashtags	Terrorism related to al-Qaeda and ISIS
#DespitetheDisbelievers	Hashtags	Refers to disbelievers of al-Qaeda and ISIS-related ideology
#IslamicState (#ISIS, #ISIL)	Hashtags	A name for al-Qaeda and ISIS
#caliphatestate	Hashtags	A name for al-Qaeda and ISIS
#Beheadings	Hashtags	Signature method for al-Qaeda and ISIS-related terrorism. It refers to the removal of the head from the body.

Table A.6.2b - Violent Right-Wing Extremism (far-right, extreme nationalistic groups & movements, neo-nazis)

Keyword	Type of keyword	Comments
Richard Spencer	Actor/Agent	American neo-Nazi, antisemitic conspiracy theorist, and white supremacist.
David Duke	Actor/Agent	American white supremacist, antisemitic conspiracy theorist, far-right politician, convicted felon, and former leader of the Ku Klux Klan.
Nikolai Nerling	Actor/Agent	German Right-Wing Extremist, anti-Semite and Holocaust denier. Alias name is "Der Volkslehrer".
David Lane	Actor/Agent	American domestic terrorist, white separatist, neo-Nazi, convicted felon, and important ideologist of white supremacy.
Olena Semenyaka	Actor/Agent	Figurehead of the Azov movement and Ukrainian Nationalism
Blood and Honour	Group/Org	International neo-Nazi/racist skinhead group started by British white supremacist
Atomwaffen Division	Group/Org	international far Right-Wing Extremist and neo-Nazi terrorist network
Nordic Resistance Movement	Group/Org	Pan-Nordic neo-Nazi movement in the Nordic countries and a political party in Sweden
Kolovrat	Group/Org	Far-right groups in a number of Slavic countries, mainly in Russia and Ukraine, use the Kolovrat in place of a swastika. It was also included in the emblem of the Russian Right-Wing radical movement Russian National Unity.
Prussian Blue	Group/Org	Prussian Blue was an American white power music duo formed in early 2003 by their mother April Gaede, a member of the neo-Nazi organization National Vanguard. Their music was described as racist and white supremacist, promoting neo-Nazi rhetoric such as Holocaust denial.
rapefugees	Phrase/Slogan	Blend of words rape and refugee.
Odinism	Phrase/Slogan	Ideology of Germanic paganism
1488	Phrase/Slogan	White Supremacist Hate Symbol. More context: The first symbol is 14, which is shorthand for the "14 Words" slogan: "We must secure the existence of our people and a future for white children." The second is 88, which stands for "Heil Hitler" (H being the 8th letter of the alphabet).
white power accelerationism	Phrase/Slogan	Idea that is inspiring white supremacist killers around the world; promotion of white supremacy
Diversity is a code word for <u>white genocide</u>	Phrase/Slogan	"Diversity = White Genocide" is a white supremacist slogan
#whitegenocide	Hashtagsf	White supremacy Ideology that white supremacy is dying
#refugeesnotwelcome	Hashtags	White supremacy idea that refugees do not belong.

#pegida	Hashtags	Patriotic Europeans Against the Islamicisation of the Occident, abbreviated Pegida, is a pan-European, anti-Islam, far-right political movement.
#istandwithrussia	Hashtags	Showcases alliance with Russia
#whitelivesmatter	Hashtags	Promotion of white supremacy

Table A.6.2c - Violent Left-Wing Extremism (anarchism, communism, etc.)

Keyword	Type of keyword	Comments
Pierre Carette	Actor/Agent	Leader of radicalised Left-Wing group (Belgium)
Otelo Saraiva de Carvalho	Actor/Agent	Leader of radicalised Left-Wing group (Portugal) - deceased
Andreas Baader	Actor/Agent	RAF terrorist
Gudrun Ensslin	Actor/Agent	RAF terrorist
Communist Combatant Cells (CCC)	Group/Org	Belgian
The Revolutionary People's Liberation Party/Front	Group/Org	Known radicalised Left-Wing group
Organisation Anarchist Action	Group/Org	Known radicalised Left-Wing group
International Revolutionary Front	Group/Org	Known radicalised Left-Wing group
Red Army Faction	Group/Org	Communist, anti-imperialist, and urban guerrilla group engaged in armed resistance against what they deemed to be a fascist state
Red Brigades	Group/Org	Militant Left-Wing organisation in Italy linked to violent acts
Lone actors	Phrase/Slogan	Violent anarchist terrorists
We shall force them to their knees	Phrase/Slogan	Antifa phrase
extreme anarcho-socialists	Phrase/Slogan	Individuals that promote violence with anarchism
Antifascist	Phrase/Slogan	A member of ANTIFA or anyone against facism
Eat the rich	Phrase/Slogan	ANTIFA slogan against capitalism
#capitalistmedia	Hashtags	Term to say that media only follows capitalist ideas and rejects anarchism
#anarchism	Hashtags	Political philosophy and movement that is sceptical of all justifications for authority and seeks to abolish the institutions
#pan-destroyer	Hashtags	Associating violence with Marxism
#antifa	Hashtags	Left-Wing anti-fascist and anti-racist political movement in the United States
#acab	Hashtags	Left-Wing acronym for All Coppers Are Bastards

6.3 External Experts

Trust Lab partnered with a number of highly qualified EU and internal experts during the data collection and labelling phase of the project, to ensure exceptional data quality and integrity.

For the data labelling process, we worked with academic TVEC experts who were also native speakers in one or several of the researched languages. As well as industry practitioners with experience in TVEC analysis and moderation, who were also proficient in the researched languages. Both groups (academic and industry), independent from each other, labelled the collected data according to our provided standards.

Trust Lab Internally, Amre Metwally, a TVEC expert and lawyer who worked at YouTube and Clubhouse for many years, oversaw the quality process. He performed regular quality checks on the work done by external parties and routinely provided feedback to improve existing processes.

6.4 Task 1: Findability Charts

6.4.1 Findability per Platform

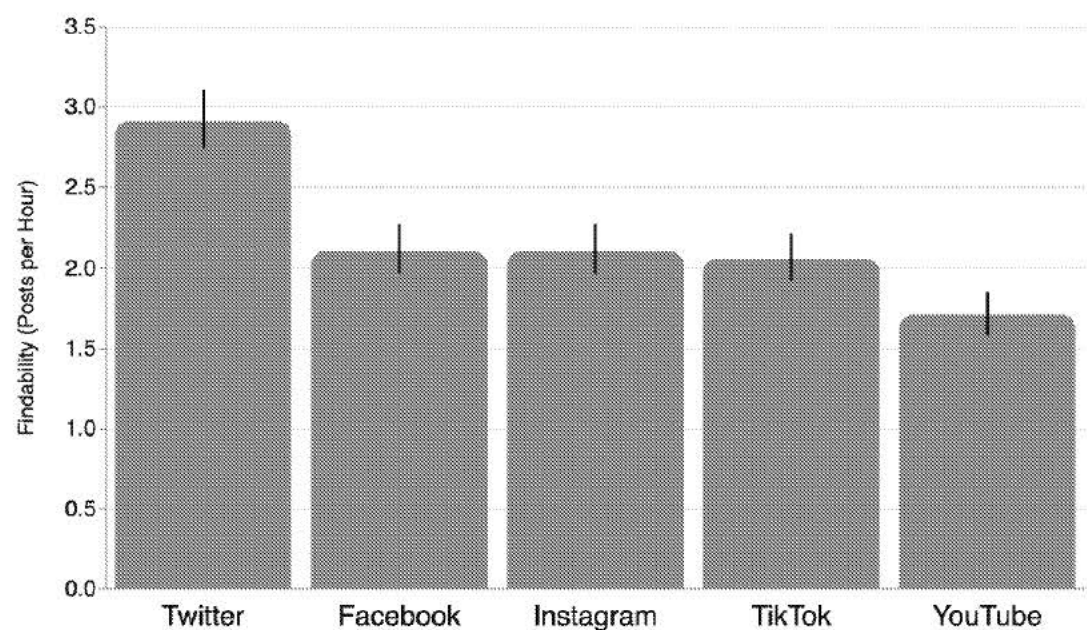


Figure A.6.4.1

The Findability Score is the average amount of content a motivated user can find on a given platform in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0012	0.0014	0.0030	0.0000
Twitter	0.0000	0.0000	0.0000	
TikTok	0.7036	0.7137		
Instagram	0.9930			

Table A.6.4.1

P-values on Findability per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.4.2 Findability per Language

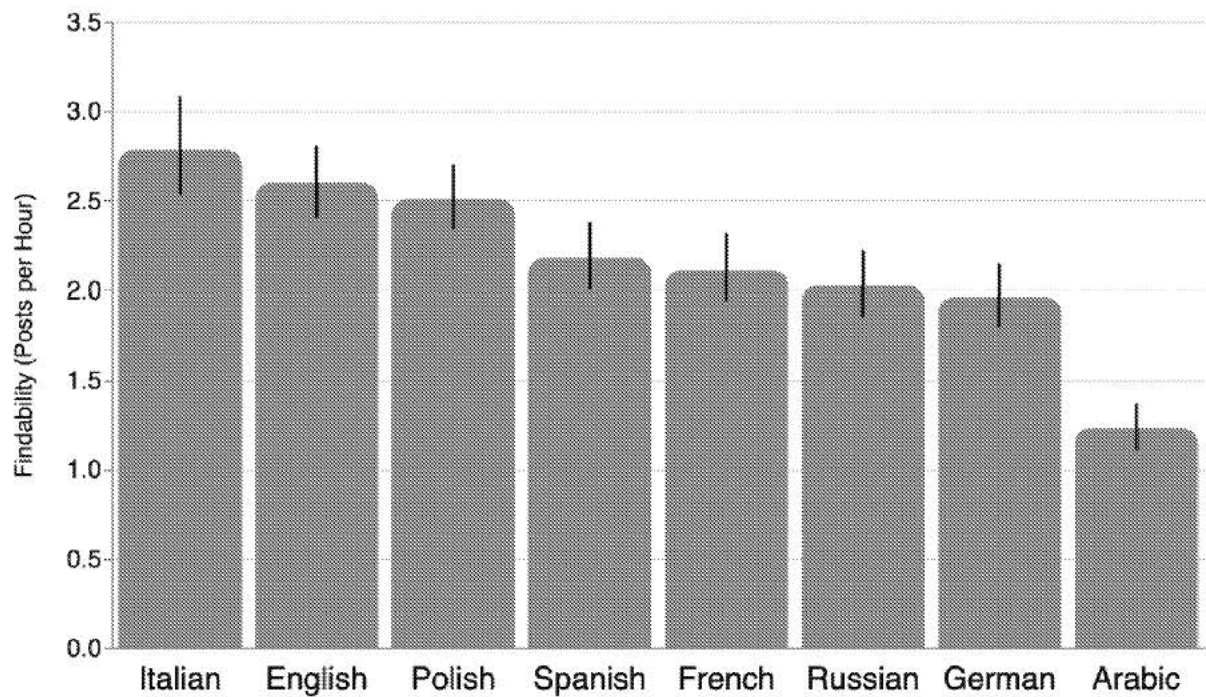


Figure A.6.4.2

The Findability Score is the average amount of content a motivated user can find in a given language in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0000	0.0115	0.6674	0.1624	0.0029	0.0348	0.3312
Russian	0.0000	0.0006	0.5919	0.6856	0.0002	0.0019	
Polish	0.0000	0.5823	0.0113	0.0003	0.1694		
Italian	0.0000	0.3710	0.0009	0.0000			
German	0.0000	0.0001	0.3426				
French	0.0000	0.0034					
English	0.0000						

Table A.6.4.2

P-values on Findability per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.4.3 Findability per TVE Type

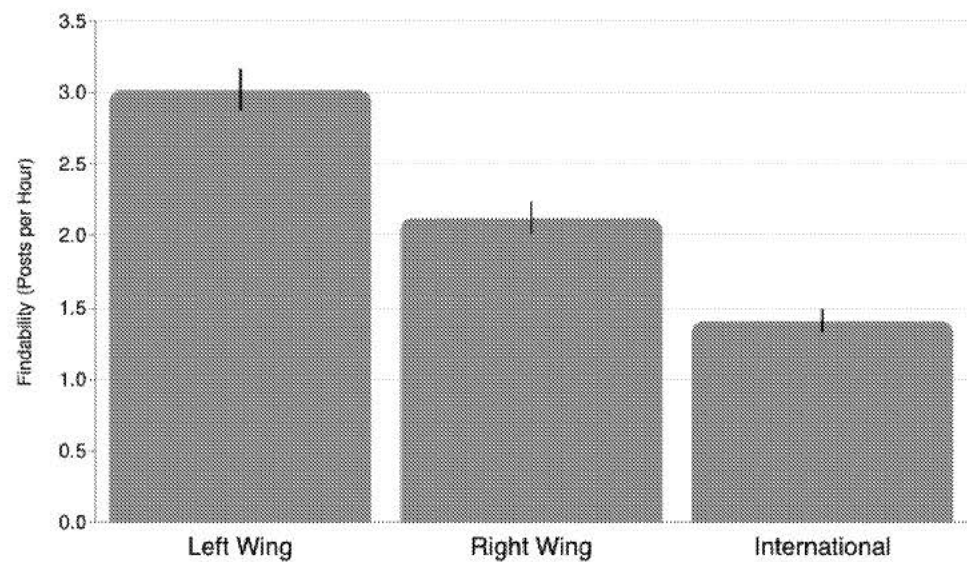


Figure A.6.4.3

The Findability Score is the average amount of content a motivated user can find for each TVE type in a one-hour search period. The block lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0000	0.0000
Left Wing	0.0000	

Table A.6.4.3

P-values on Findability per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.4.4 Findability per Language / TVE Type

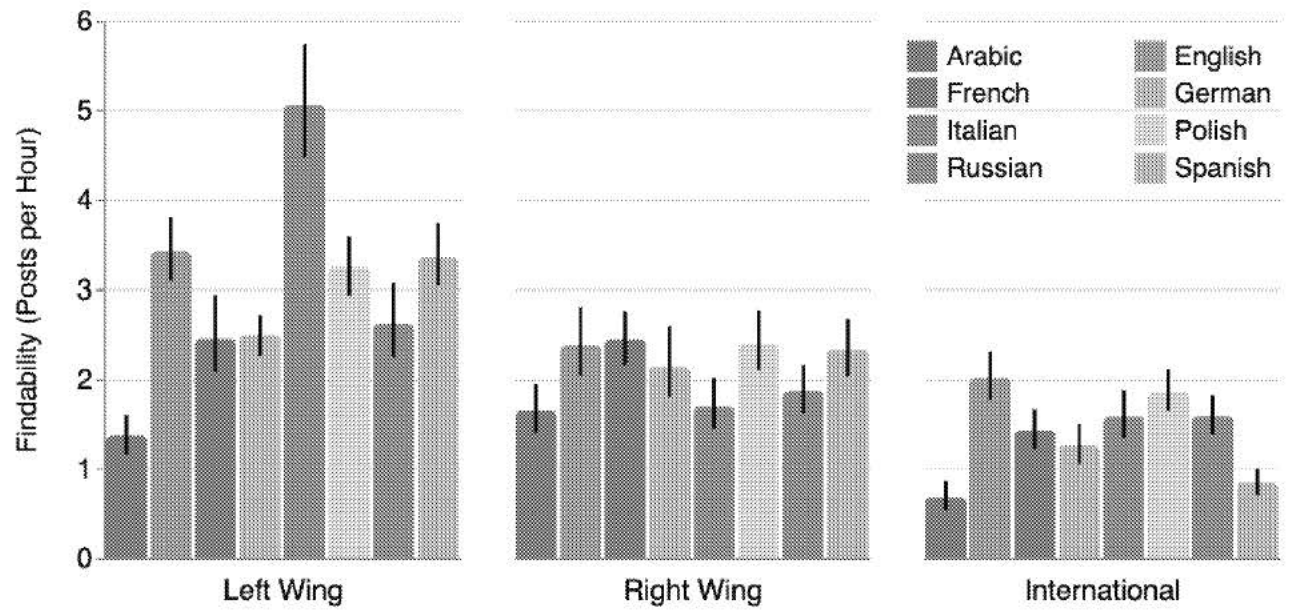


Figure A.6.4.4

The Findability Score is the average amount of content a motivated user can find for the respective language and TVE type in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

6.5 Task 1: Examples of Italian Violent Left-Wing Extremist Content

To help illustrate some of the Violent Left-Wing Extremist content, here is an example of a relatively benign [video](#) on YouTube that supports anarchism:



Image A.6.5a

The video title translates to “What if you were also an Anarchist?” and the bio of the user reads (translated using Google Translate):

“Anyone who hears the term Anarchy thinks of chaos, disorder, violence. Family upbringing and society have inculcated a totally false and distorted view of Anarchy into people’s minds. Try doing this kind of test and find out if, like me, you are an Anarchist.”

- A more extreme example is below (warning: violent), where the poster of the [Tweet](#) has the name “Antifa_Ultras” and includes the hashtag ACAB (All Cops are Bastards) accompanying a photo of uniformed police officers, one who is surrounded by fire. The name of the user (Antifa_Ultras) or the hashtag alone, without any other signal, might indicate Borderline content by indicating association, or at the very least affinity for, a Left-Wing organisation (Antifa) or a Left-Wing ideology (anti-police). However, with the image, the content, in the eyes of the outside expert, would constitute a violation for encouraging the use of violence:



Image A.6.5b

6.6 Task 1: Examples Twitter Content

Some examples of content that was discovered on Twitter. This section starts with Right-Wing (Borderline) content, then Left-Wing (Borderline) content, then concludes with International (Borderline) content.

6.6.1 Violent Right-Wing Extremism and Borderline Examples

- Borderline anti-immigrant and anti-refugee rhetoric:



Image A.6.6 1a

- “White lives matter” rhetoric from anti-immigrant voices in Europe:



Image A.6.6.1b

- Praise for Richard Spencer and other leading figures and voices in the white supremacist movement. The user's bio purports that the account is intended to be a parody. While many

platforms may have exceptions for satirical or comedic purposes in certain instances, they become difficult to enforce. One key reason is that discrete pieces of content often lack broader context, especially if platform review processes separate user-level review from content review. If so, then a reviewer simply looking at this screenshot would see clear praise for Richard Spencer based on the text. If reviewers are able to see this piece of content along with the user bio that states it is a parody account, the analysis becomes more complicated.



Image A.6.6.1c

6.6.2 Violent Left-Wing Extremism and Borderline Examples

- Critiques of capitalism and capitalist structures (e.g., media):

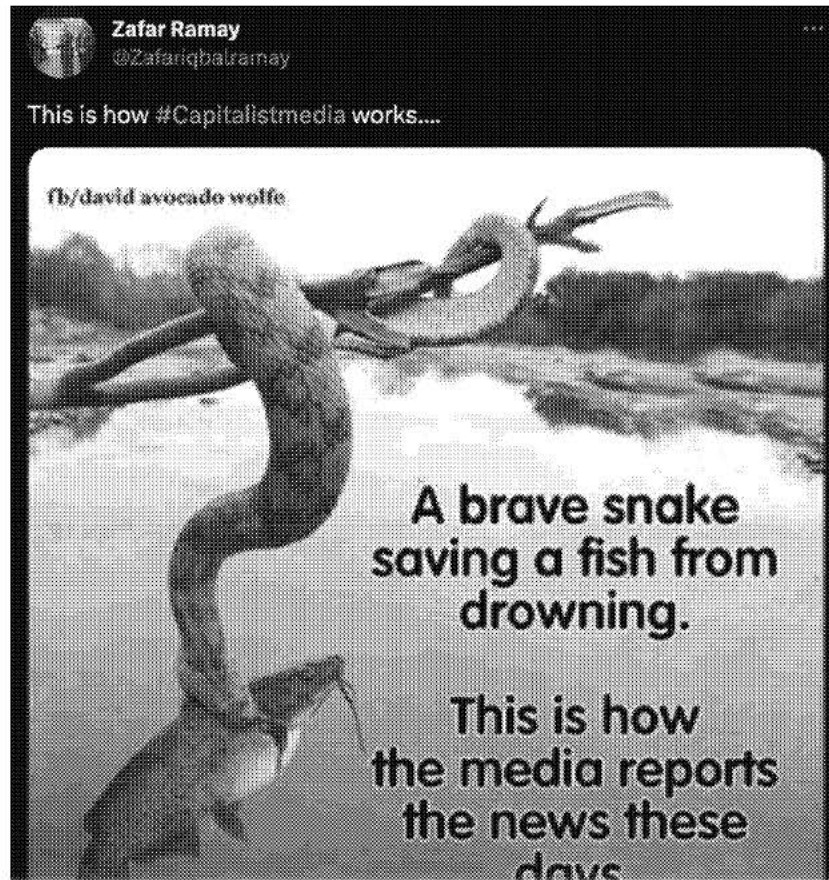


Image A.6.6.2a

- Violent Left-Wing Extremist content exemplified through anarchist rhetoric :



Image A.6.6.2b

- Anti-police commentary:



Image A.6.6.2c

6.6.3 International Extremism and Borderline Examples

Of the content related to Islamist TVE, it seemed to focus primarily on ISIS, Hezbollah, al-Qaeda, and Hizb-ut-Tahrir material. Examples include:

- Accounts dedicated to Hizb-ut-Tahrir "branches" in specific countries. The text says: "Independence of the flag from the colonialists".

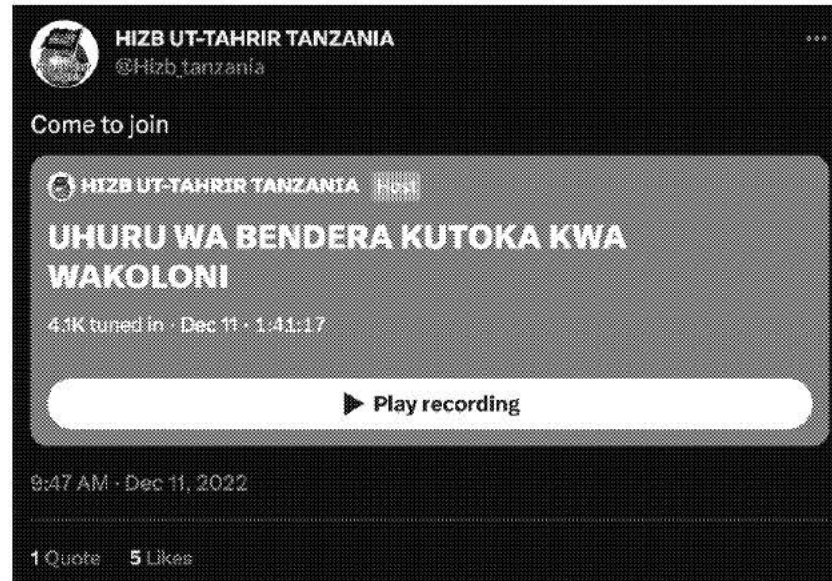


Image A.6.6.3a

- Praise for al-Qaeda leaders:

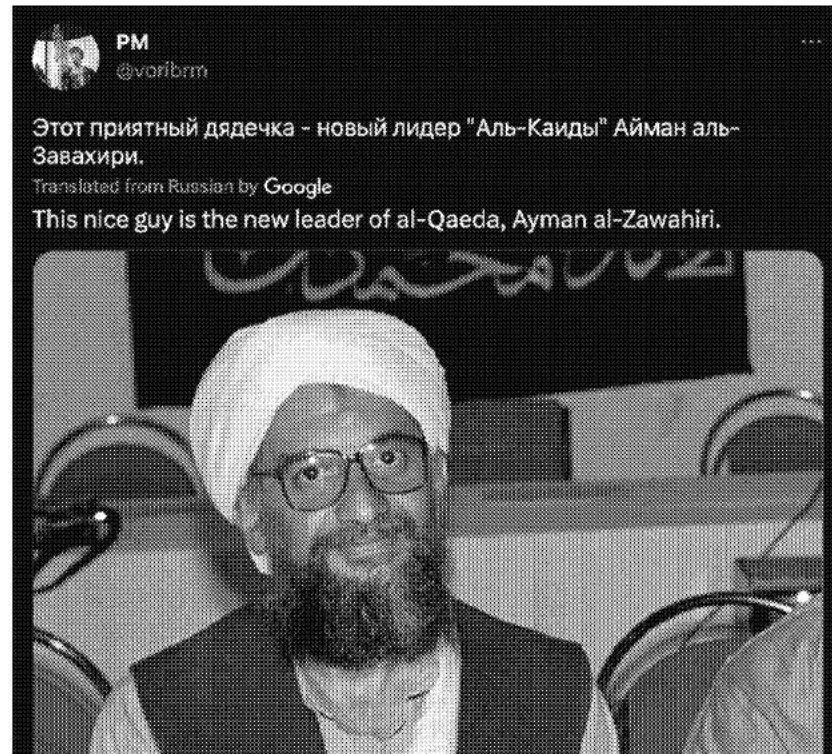


Image A.6.6.3b

- ISIS-produced propaganda:



image A.6.6.3c

6.7 Task 1: User Engagement Charts and P-Values

6.7.1 Average Number of Shares

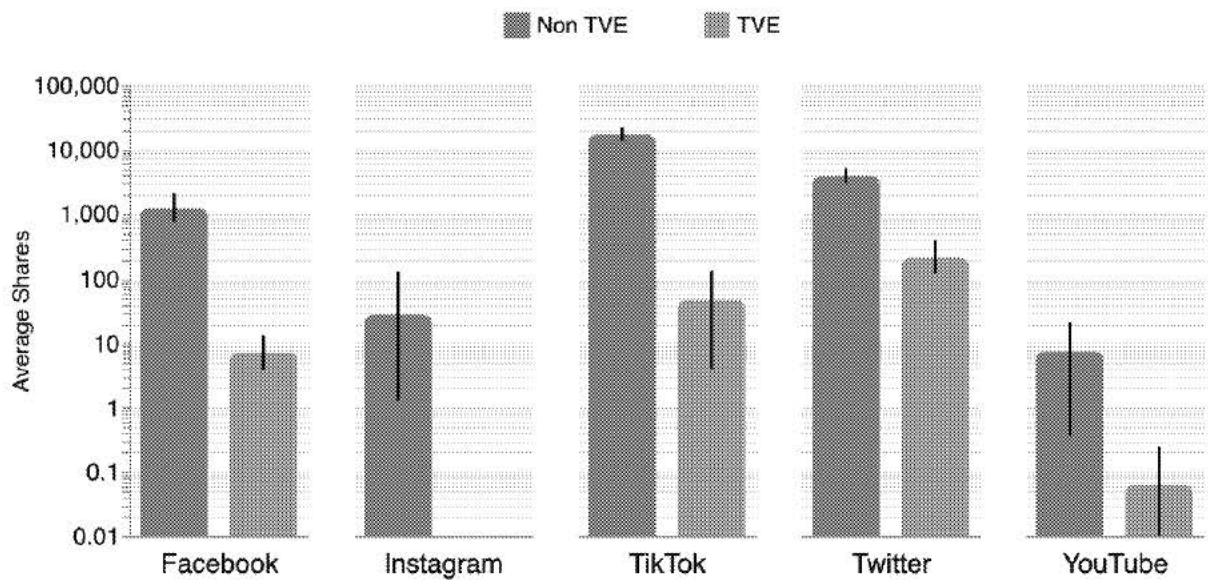


Figure A.6.7.1

This chart shows the average number of shares for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.²⁹

²⁹ Since the range of values in these charts is very high, it would be difficult to visualise the bars that have lower values. Taking the logarithm makes the values closer in range on the log scale and hence easier to visualise the chart as a whole. Therefore, for the interaction charts, we are using log scales.

6.7.2 Average Number of Likes

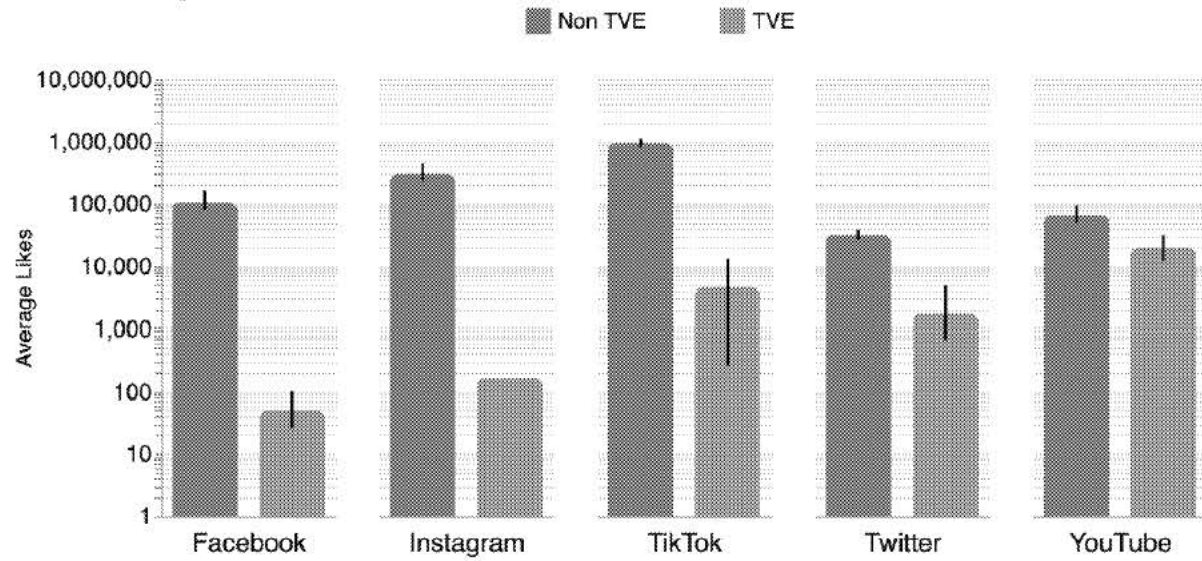


Figure A.6.7.2

This chart shows the average number of likes for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.

6.7.3 Average Number of Comments

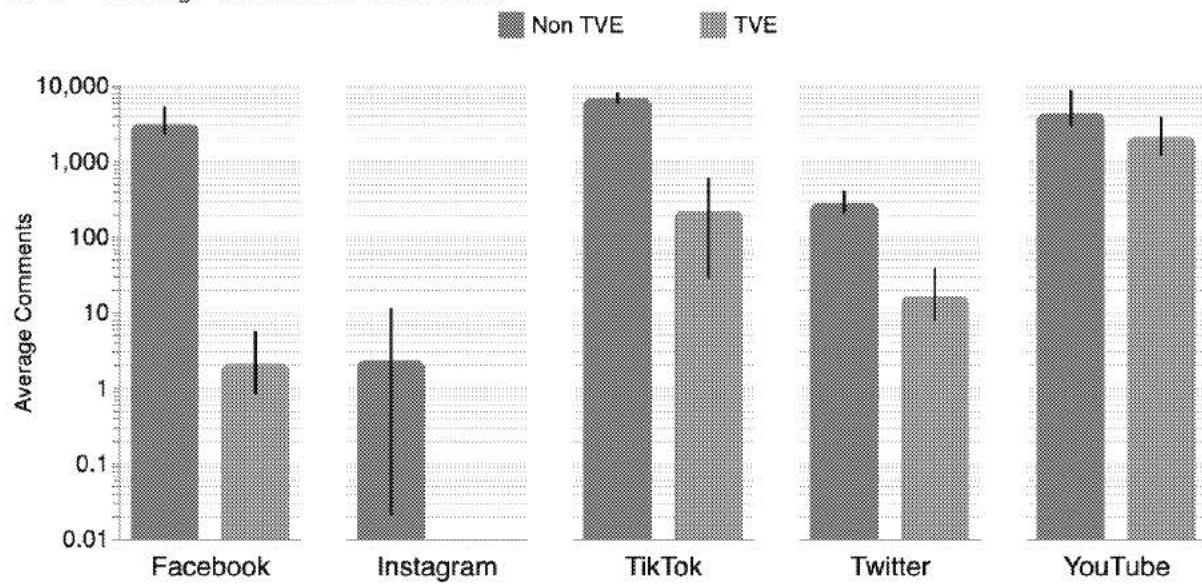


Figure A.6.7.3

This chart shows the average number of comments for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the interaction charts all use a log scale instead of a linear scale.

6.7.4 Average Number of Followers

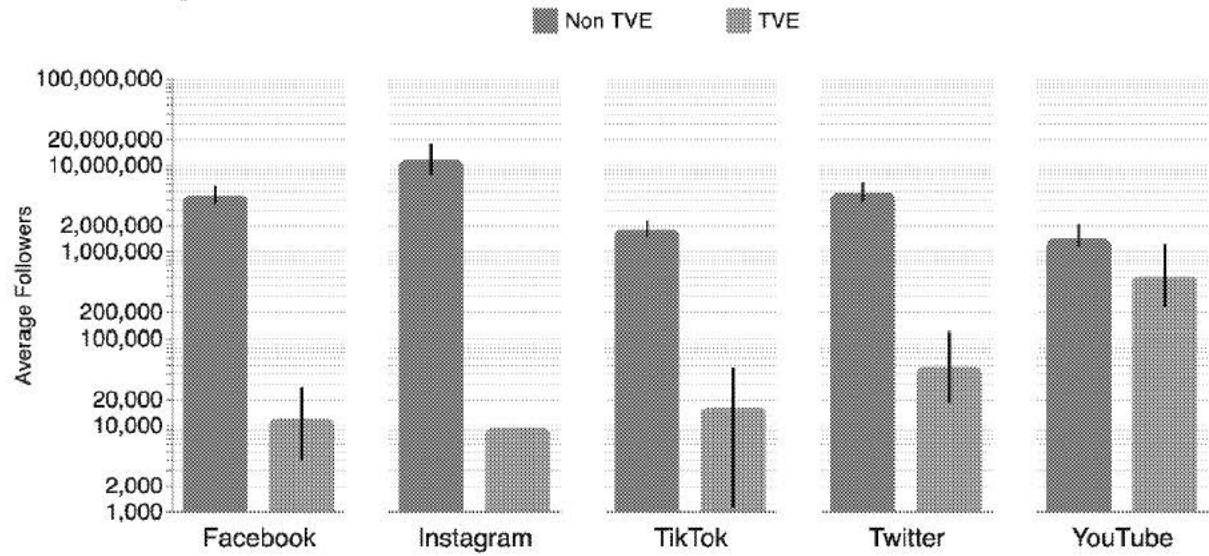


Figure A.6.7.4

This chart shows the average number of followers for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.

6.8 Task 1: Removal Rates for TVE Content

6.8.1 Removal Rates per Platform

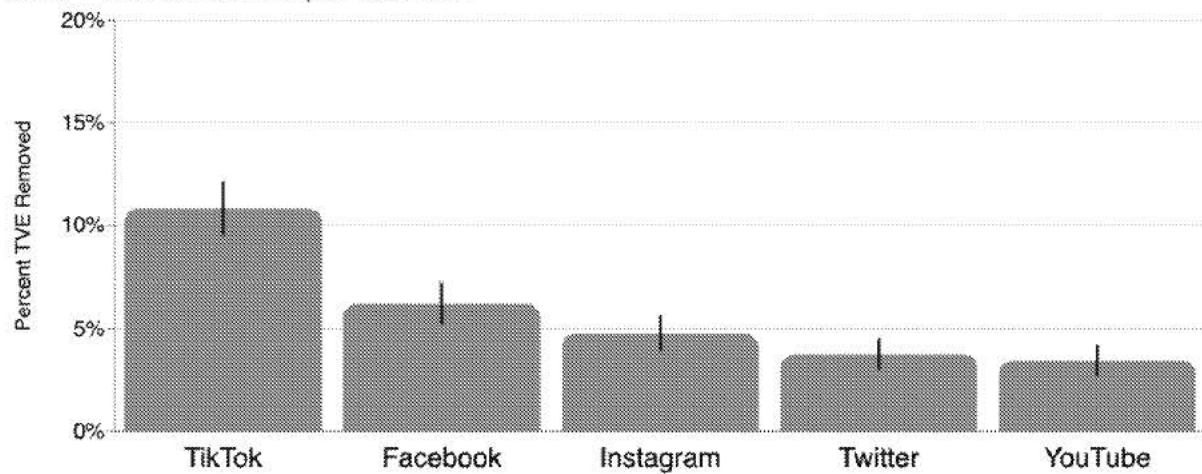


Figure A.6.8.1
This graph shows the percentage of TVE content that was removed on a given platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0311	0.2628	0.0000	0.7646
Twitter	0.0497	0.3870	0.0000	
TikTok	0.0062	0.0002		
Instagram	0.2802			

Table A.6.8.1
P-values on Removal Rates per Platform
alpha = 0.01 (Bonferroni correction from 0.05)

6.8.2 Removal Rates per Language

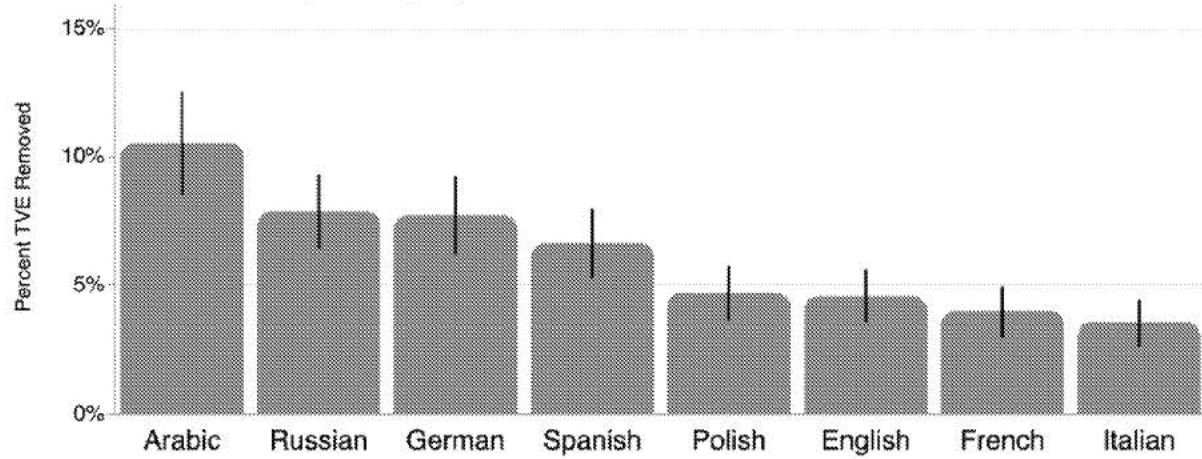


Figure A.6.8.2

This graph shows the percentage of TVE content that was removed in a given language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0957	0.2222	0.1053	0.5905	0.0490	0.2588	0.5354
Russian	0.2750	0.0618	0.0234	0.9472	0.0086	0.0766	
Polish	0.0053	0.9317	0.6224	0.0965	0.4042		
Italian	0.0003	0.4522	0.7349	0.0124			
German	0.2606	0.0791	0.0316				
French	0.0011	0.6825					
English	0.0039						

Table A.6.8.2

P-values on Removal Rates per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.8.3 Removal Rates per TVE Type

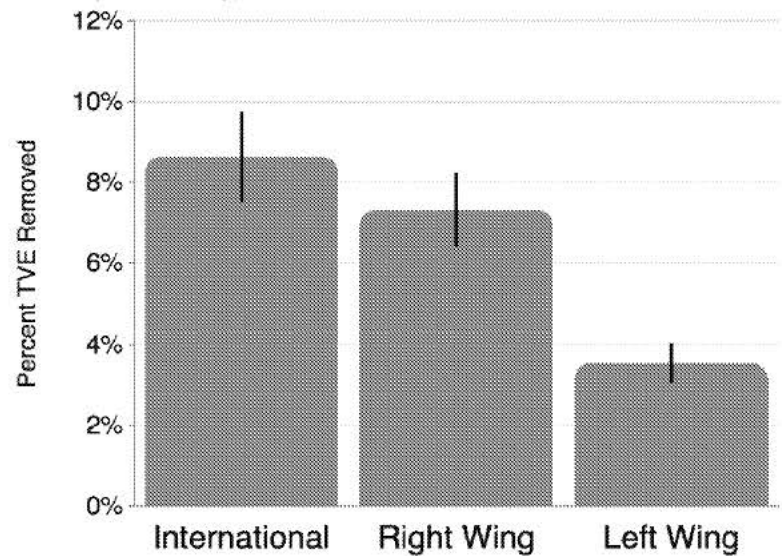


Figure A.6.8.3

This graph shows the percentage of TVE content that was removed per TVE type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.3610	0.0001
Left Wing	0.0000	

Table A.6.8.3

P-values on Removal Rates per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.8.4 Average number of Shares associated with Removed TVE

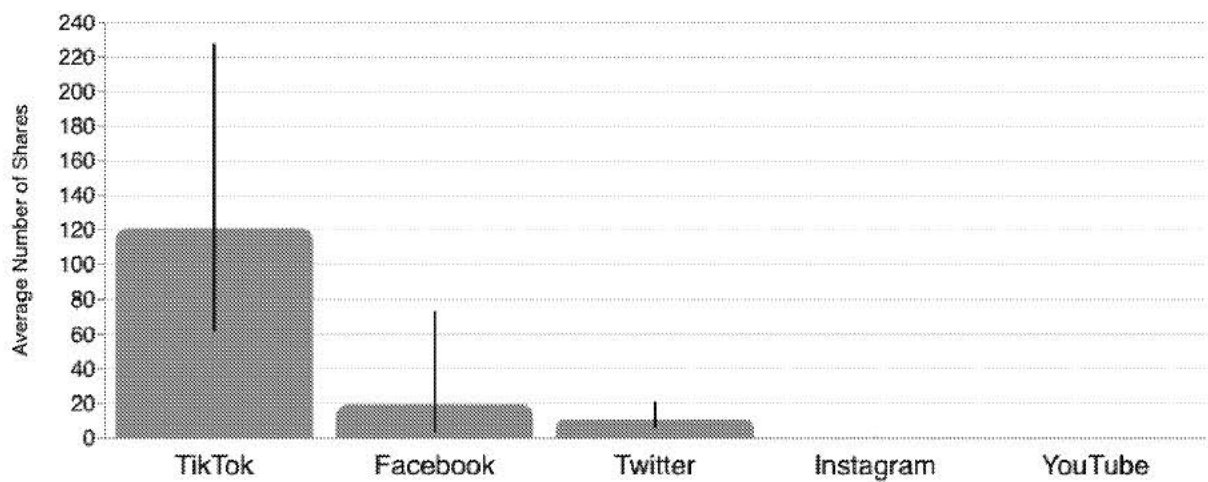


Figure A.6.8.4

This is the average number of shares associated with TVE content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.3076	0.2936	0.1426	0.0386
Twitter	0.5656	0.0160	0.1352	
TikTok	0.1001	0.0781		
Instagram	0.2784			

Table A.6.8.4

P-values on Average number of Shares associated with Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.5 Average number of Shares associated with TVE that wasn't Removed

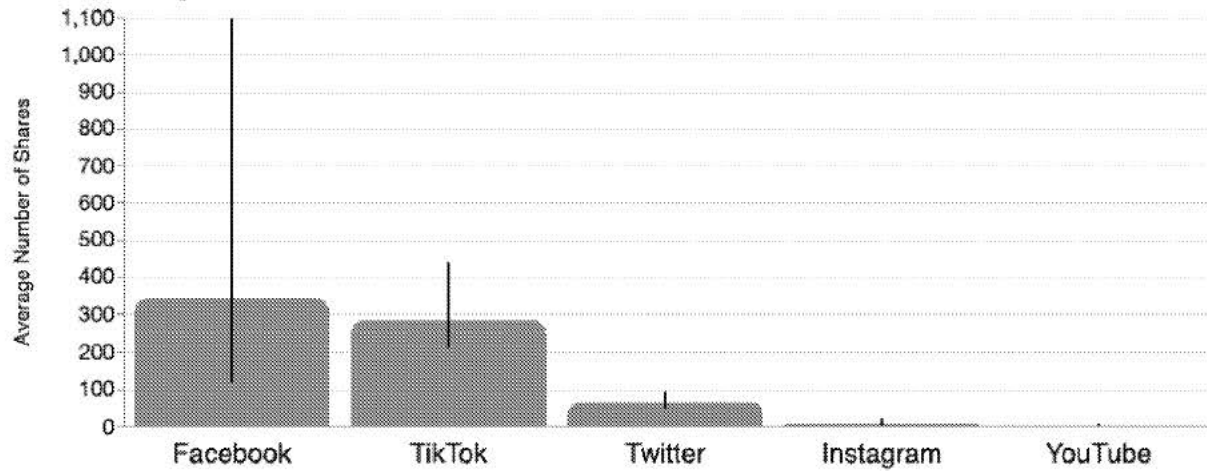


Figure A.6.8.5

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study. The block lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1245	0.3010	0.0000	0.0000
Twitter	0.1781	0.0001	0.0001	
TikTok	0.8029	0.0000		
Instagram	0.1198			

Table A.6.8.5

P-values on Average number of Shares associated with non-Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.6 Average number of Shares associated with TVE that wasn't Removed broken down by language

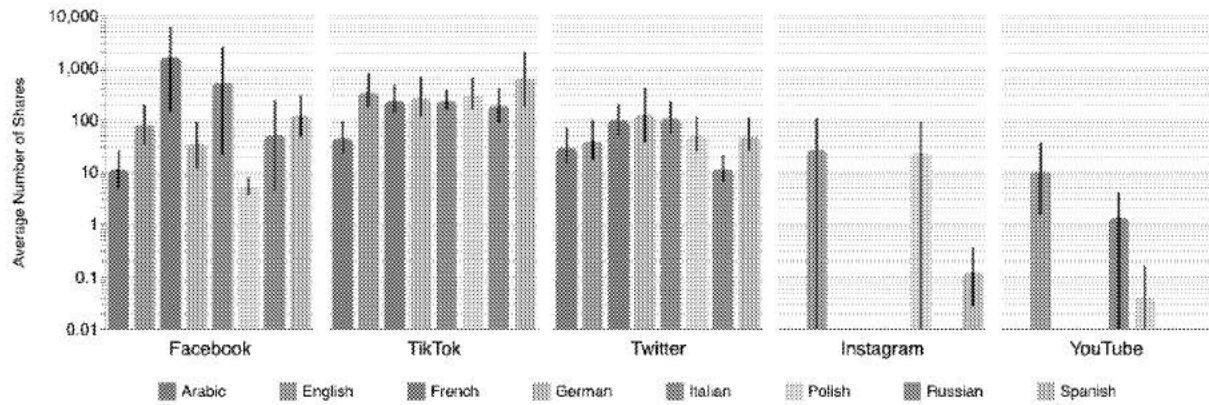


Figure A.6.8.6

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.8.7 Average number of Shares associated with Removed Borderline content

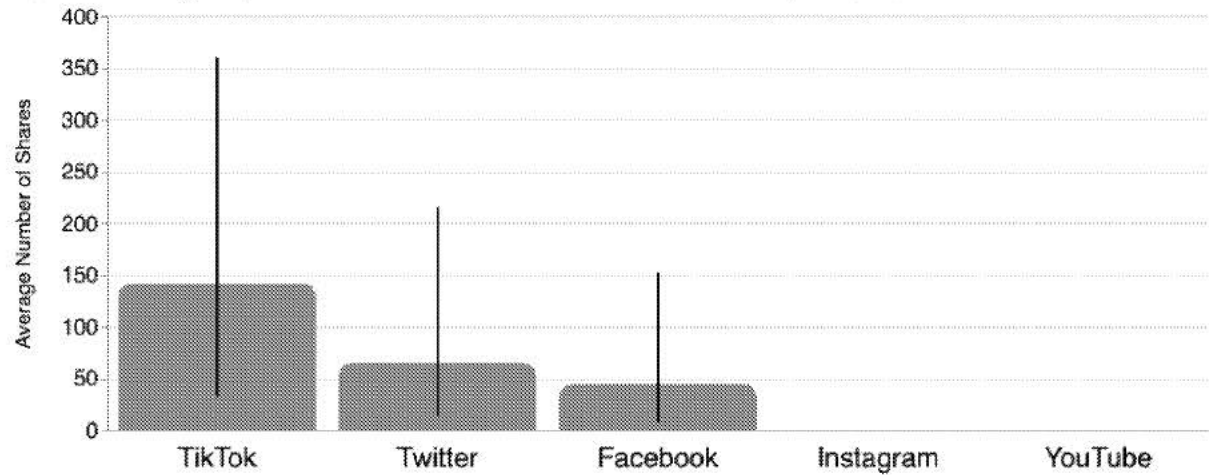


Figure A.6.8.7

This is the average number of shares associated with Borderline content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1305	0.0000	0.1387	0.1215
Twitter	0.6911	0.1194	0.4868	
TikTok	0.3310	0.1567		
Instagram	0.1596			

Table A.6.8.7

*P-values on Average number of Shares associated with Removed Borderline content
alpha = 0.01 (Bonferroni correction from 0.05)*

6.8.8 Average number of Shares associated with Borderline content that wasn't Removed

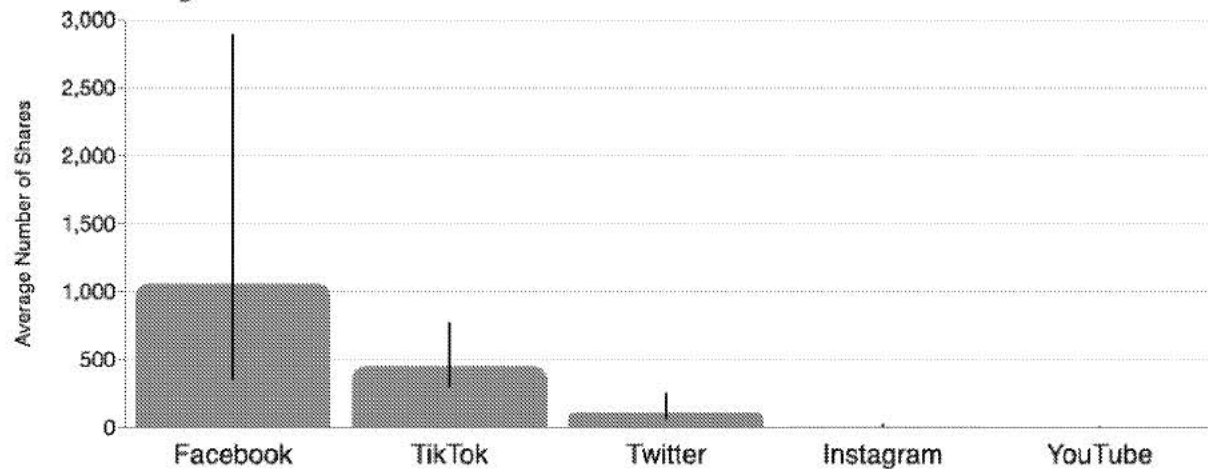


Figure A.6.8.8

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0617	0.5061	0.0001	0.0149
Twitter	0.1141	0.0302	0.0083	
TikTok	0.3682	0.0002		
Instagram	0.0861			

Table A.6.8.8

*P-values on Average number of Shares associated with non-Removed Borderline content
alpha = 0.01 (Bonferroni correction from 0.05)*

6.8.9 Average number of Shares associated with Borderline content that wasn't Removed broken down by language

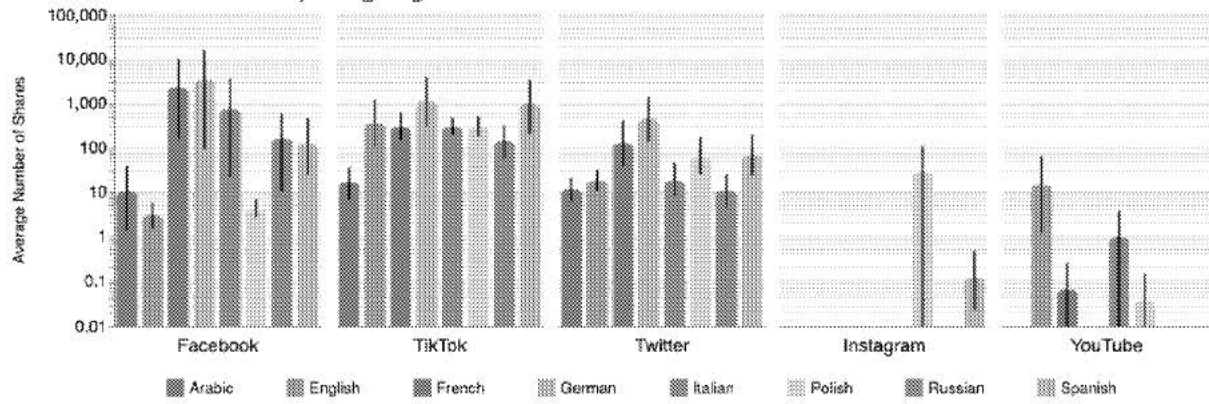


Figure A.6.8.9

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.9 Task 1: Removal Time

6.9.1 Removal Time per Platform

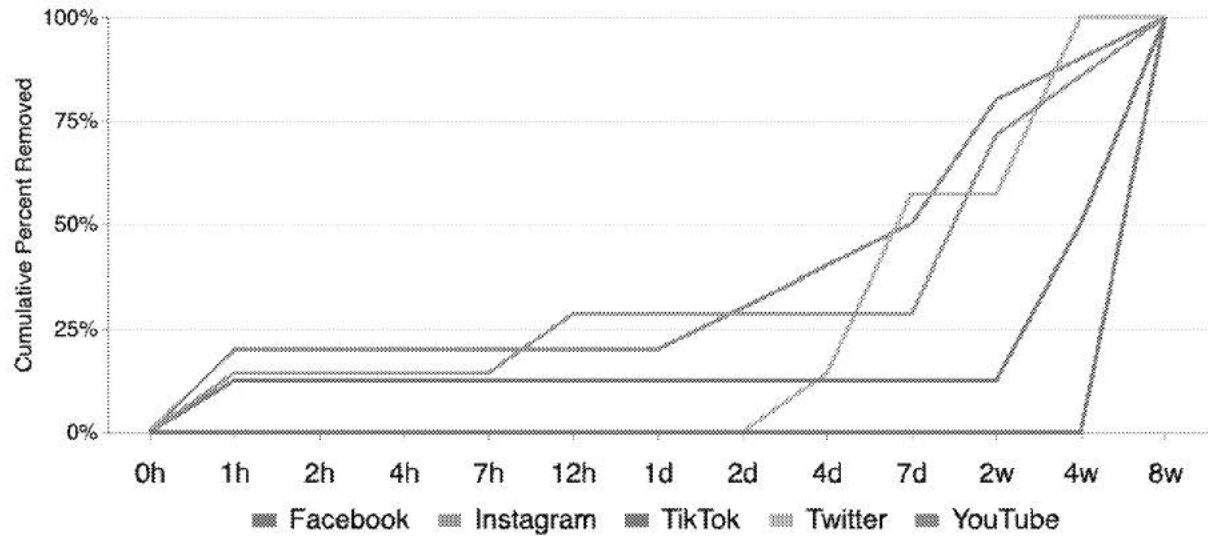


Figure A.6.9.1

This chart shows how long it took for each platform to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.2 Removal Time per Language

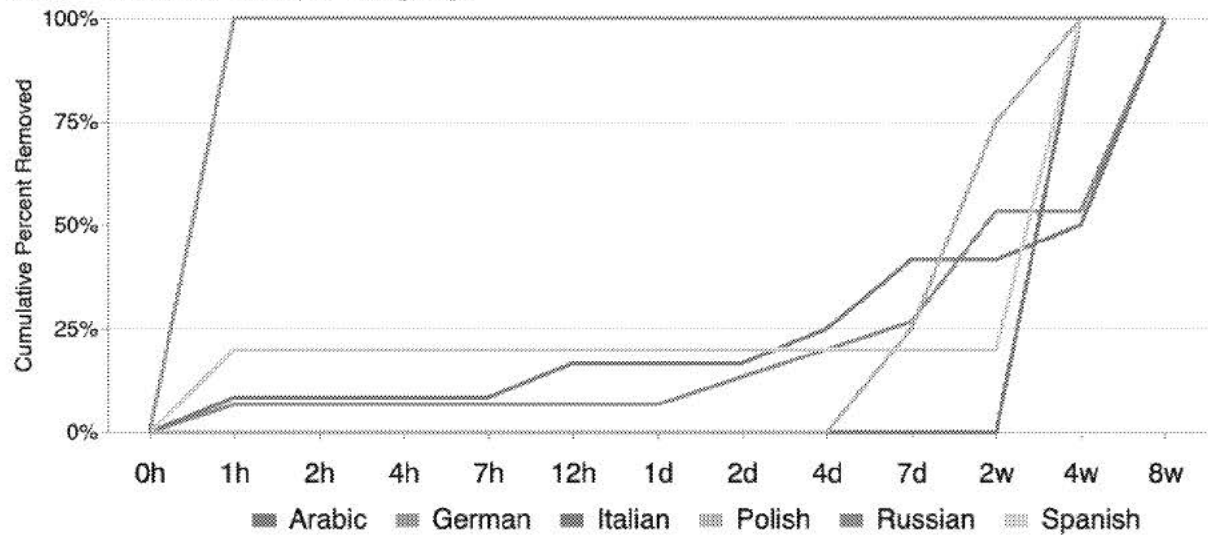


Figure A.6.9.2

This chart shows how long it took for each language to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.3 Removal Time per TVE Type

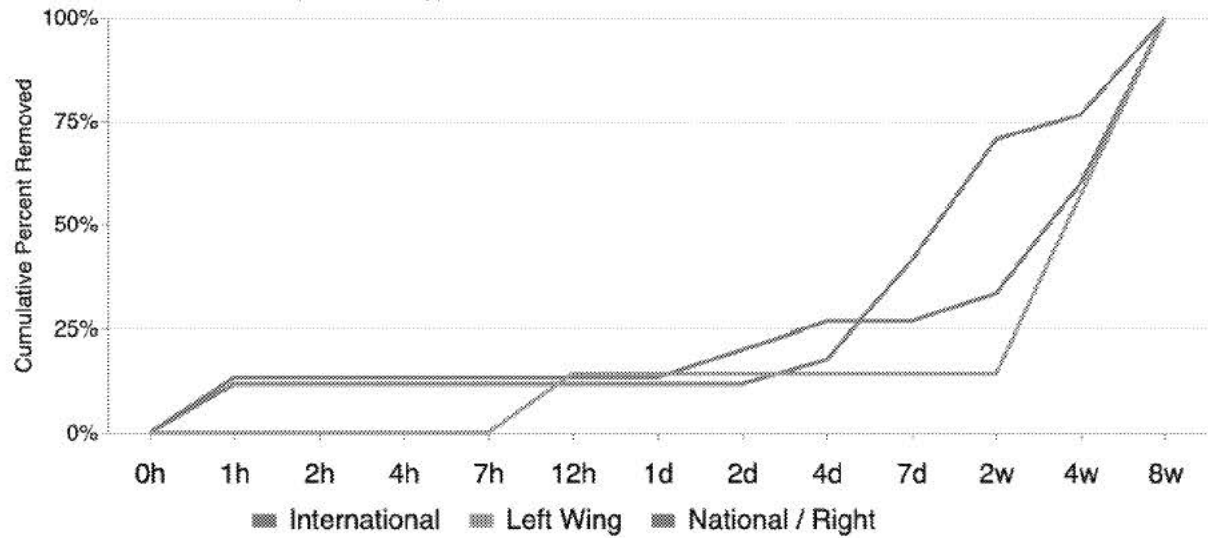


Figure A.6.9.3

This chart shows how long it took for each TVE Type to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.10 Task 1: User Sentiment Metrics

6.10.1 Severity Ratings by Platform

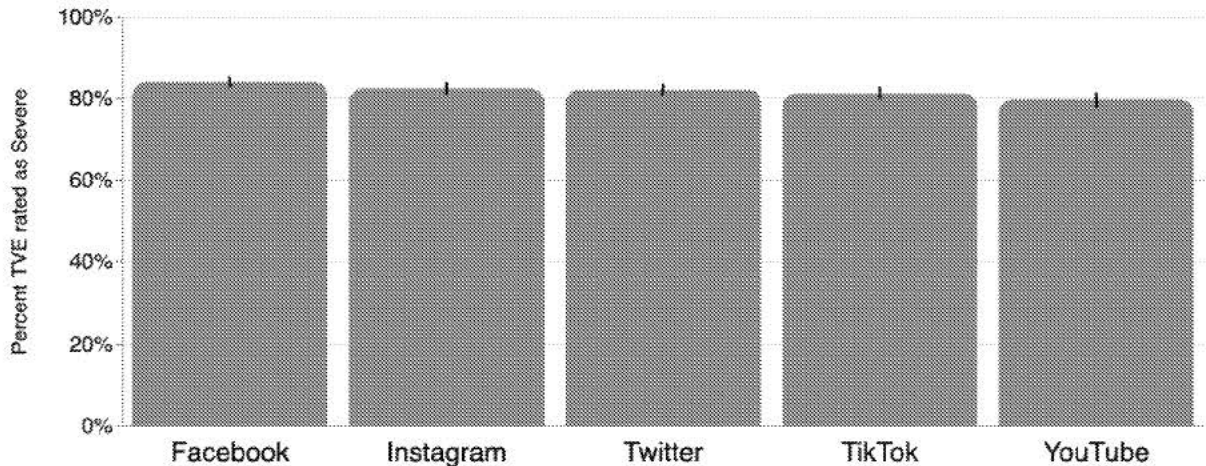


Figure A.6.10.1

This chart shows the percentage of TVE content present on different platforms that was rated by users as “severe”. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.10.1

P-values on Severity Ratings per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.10.2 Severity Ratings by Language

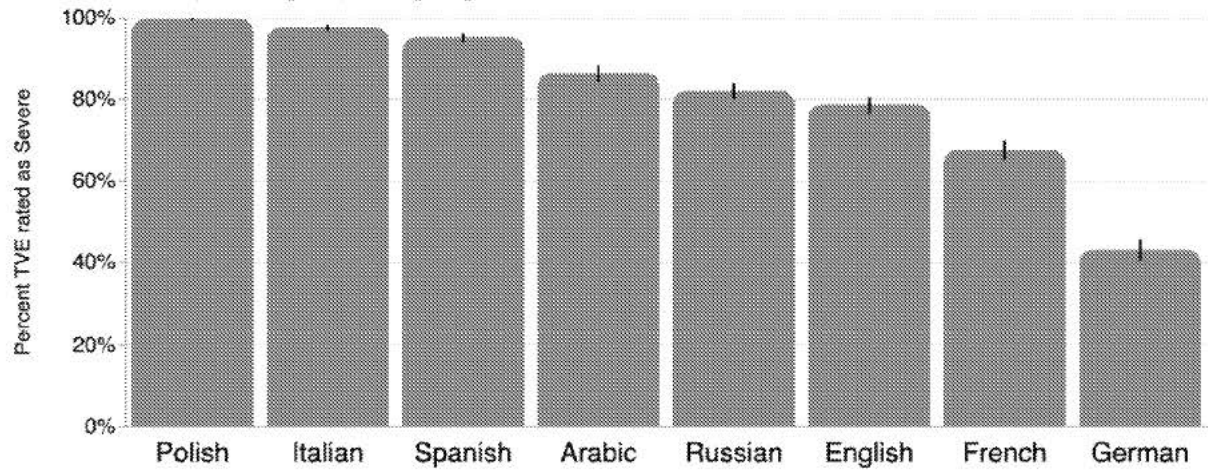


Figure A.6.10.2

This chart shows the percentage of TVE content present in different languages that was rated by users as severe. The black lines in the bars represent 90% confidence intervals.

Polish values are biased here because of the presence of bad actors in the dataset, which caused users to rate everything in the dataset as “severe”. This problem was investigated and deemed not present for the other languages.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0001	0.0000	0.0000	0.0000	0.0573	0.0000	0.0000
Russian	0.1311	0.2206	0.0000	0.0000	0.0000	0.0000	
Polish	0.0000	0.0000	0.0000	0.0000	0.0124		
Italian	0.0000	0.0000	0.0000	0.0000			
German	0.0000	0.0000	0.0000				
French	0.0000	0.0003					
English	0.0088						

Table A.6.10.2

P-values on Severity Ratings per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.10.3 Severity Ratings by TVE Type

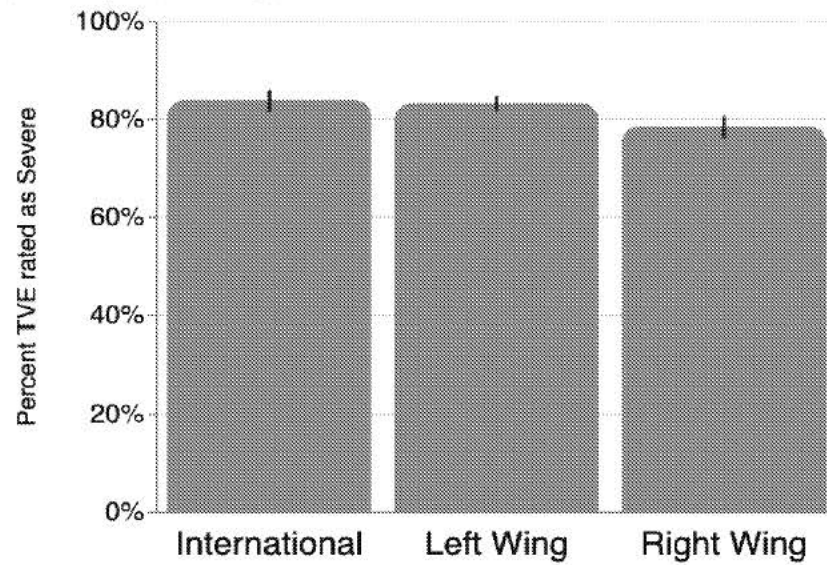


Figure A.6.10.3

This chart shows the percentage of TVE content present for different TVE Types that was rated by users as "severe". The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0074	0.0045
Left Wing	0.6996	

Table A.6.10.3

P-values on Severity Ratings per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.10.4 Severity Ratings over Time per Platform

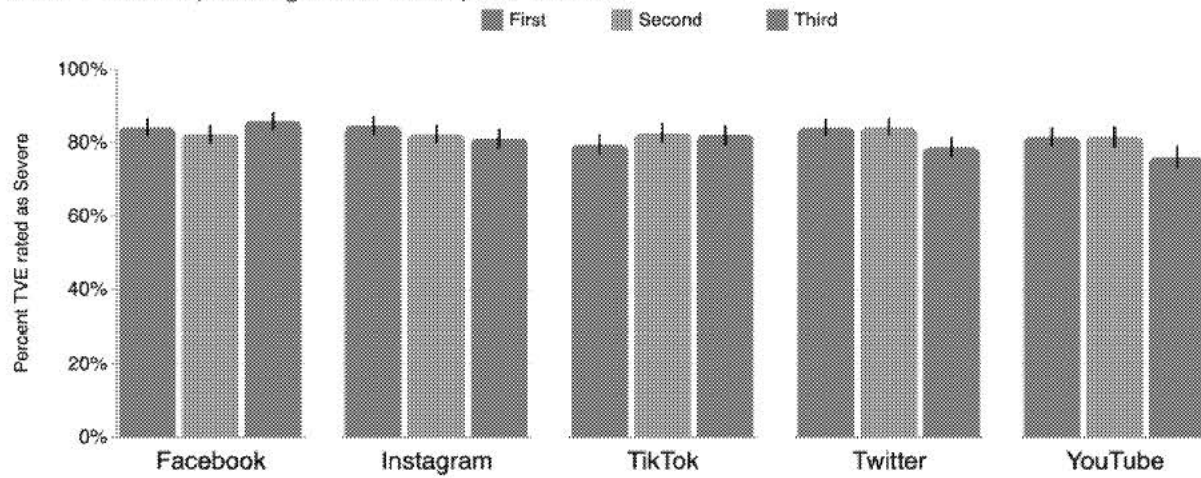


Figure A6.10.4

This chart shows the percentage of TVE content rated as "Severe". First, second and third stages are shown for each platform. The black lines in the bars represent 90% confidence intervals.

P-values are not included for this chart due to the number of combinations and complexity of the comparison.

6.11 Task 2: Amplification Charts

6.11.1 Amplification Across all Platforms

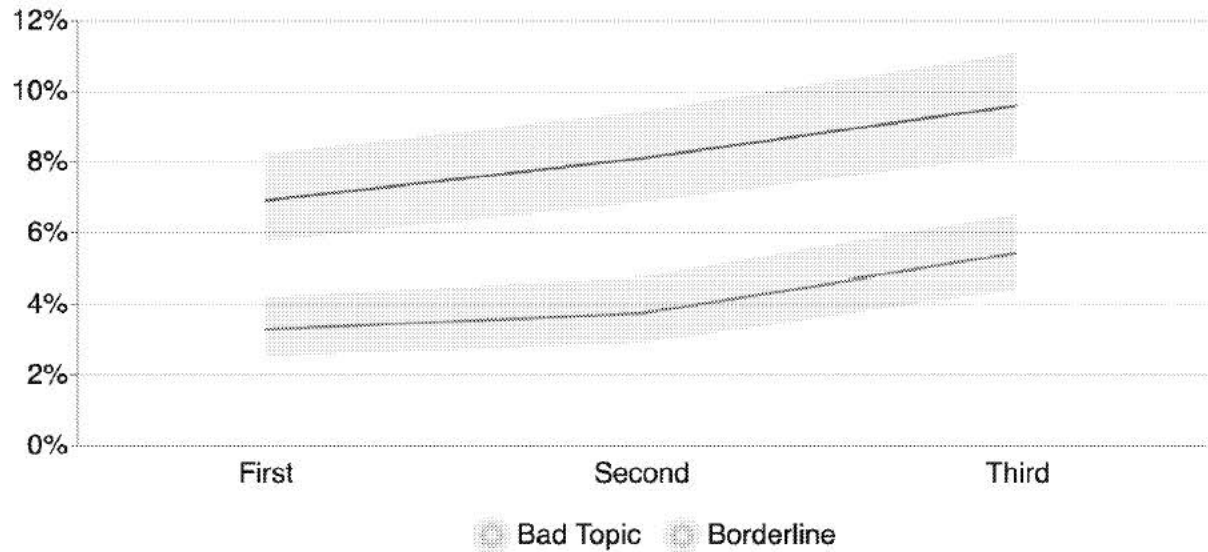


Figure A.6.11.1a

This graph illustrates the mean percentage of “Bad Topic” and “Borderline” content recommended in user feeds over time. The shaded area represents the 90% confidence interval (alpha = 0.05)

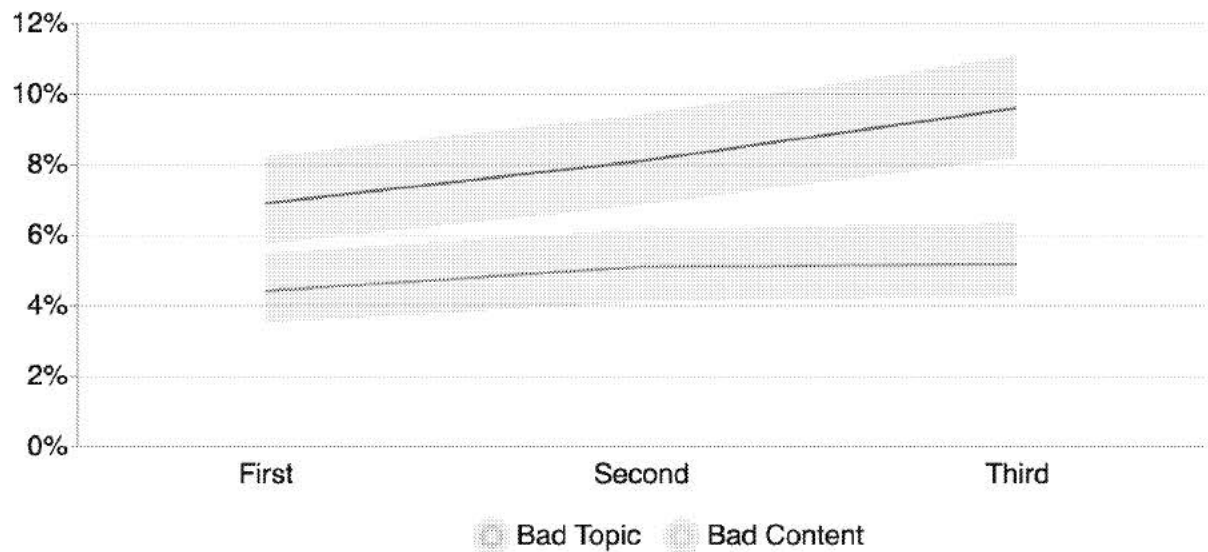


Figure A.6.11.1b

This graph illustrates the mean percentage of “Bad Topic” and “Bad Content” content recommended in user feeds over time. The shaded area represents the 90% confidence interval (alpha = 0.05)

6.11.2 Amplification Average Percent of Bad Content in Feed per Platform

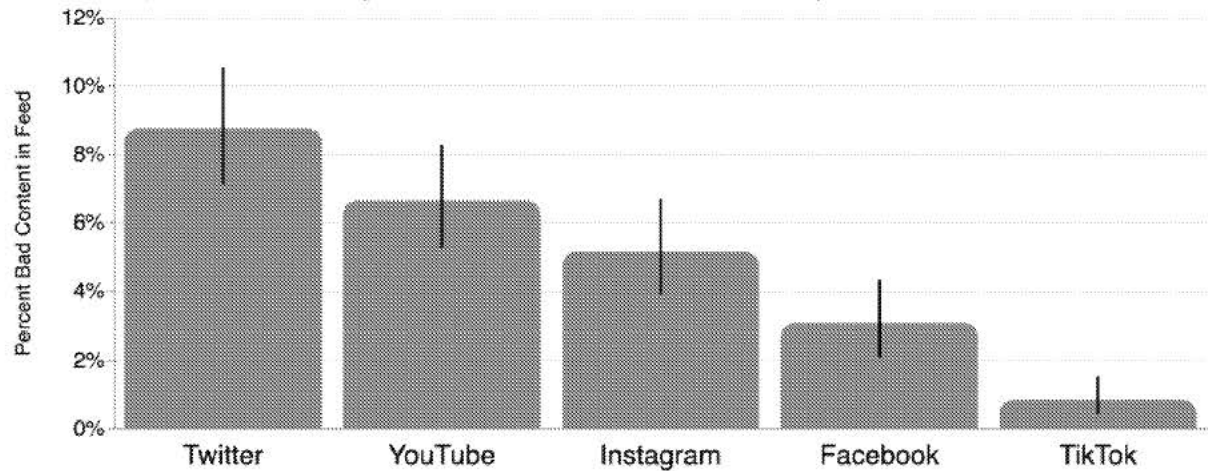


Figure A.6.11.2

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective platforms. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.11.2

P-values on Amplification Average Percent of Bad Content in Feed per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.11.3 Amplification Average Percent of Bad Content in Feed per Language

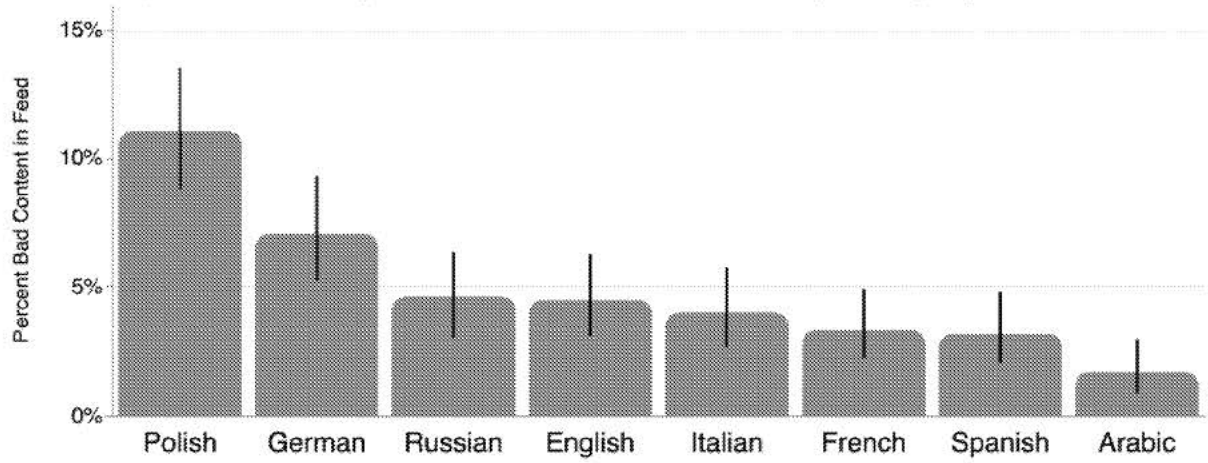


Figure A.6.11.3

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective languages. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.1466	0.3265	0.9038	0.0097	0.5335	0.0000	0.2809
Russian	0.0118	0.9246	0.3343	0.1168	0.6438	0.0003	
Polish	0.0000	0.0002	0.0000	0.0387	0.0001		
Italian	0.0381	0.7145	0.6132	0.0440			
German	0.0001	0.0986	0.0126				
French	0.1146	0.3855					
English	0.0155						

Table A.6.11.3

P-values on Amplification Average Percentage of Bad Content in Feed per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.11.4 Amplification Average Percent of Bad Content in Feed per TVE Type

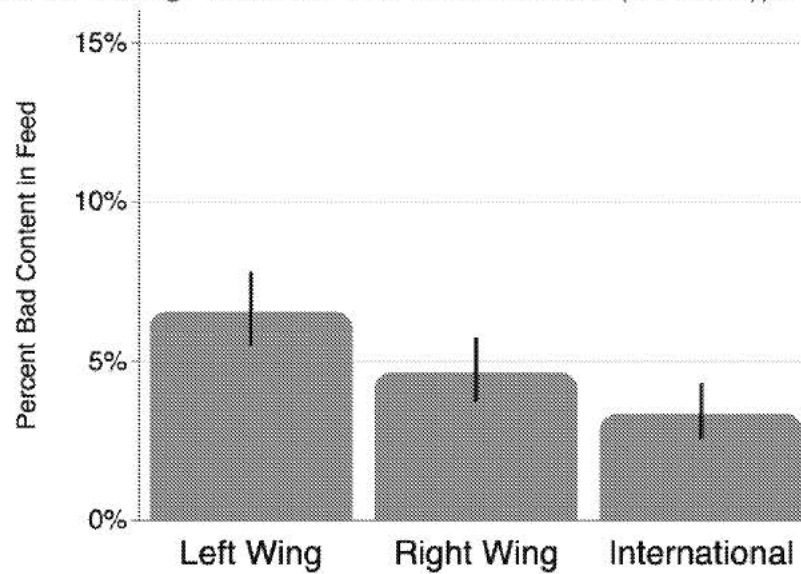


Figure A.6.11.4

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective TVE Types. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.1058	0.0442
Left Wing	0.0003	

Table A.6.11.4

P-values on Amplification Average Percent of Bad Content in Feed per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.11.5 Amplification Percentage Change for Bad Content per Content Type

To better compare data across other dimensions (platform, language, etc), we will use the percent difference between the first and third stage's mean. We don't compare to zero state because the assumption is that a new account's feed will be empty or contain zero TVE related content.

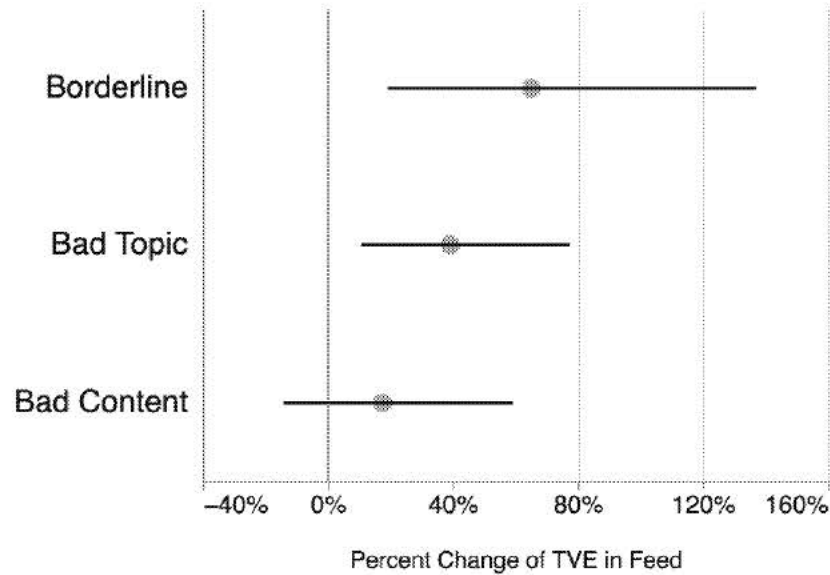


Figure A.6.11.5

This chart shows the percentage change for the amount of Bad (TVE) Content per Content Type in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.6 Amplification Percent Change for Bad Content per Platform

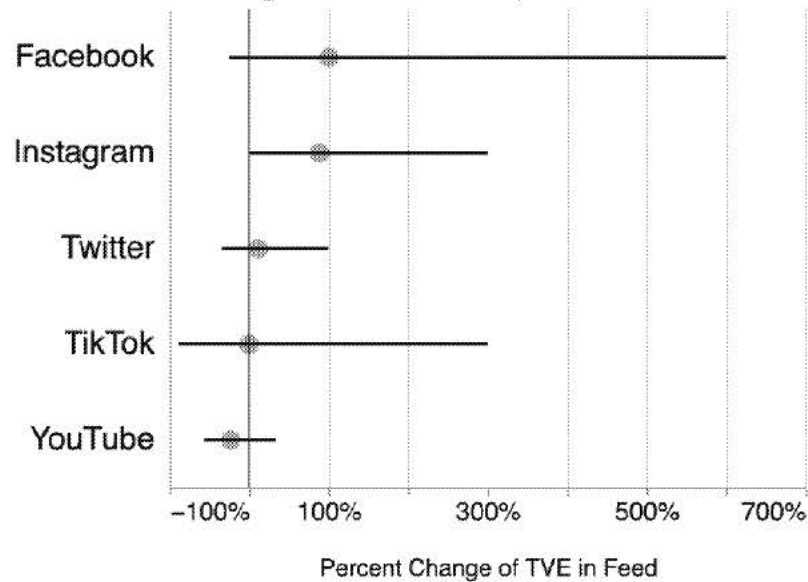


Figure A.6.11.6

This chart shows the percentage change for the amount of Bad (TVE) Content in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.7 Amplification Percent Change for Bad Content per Language

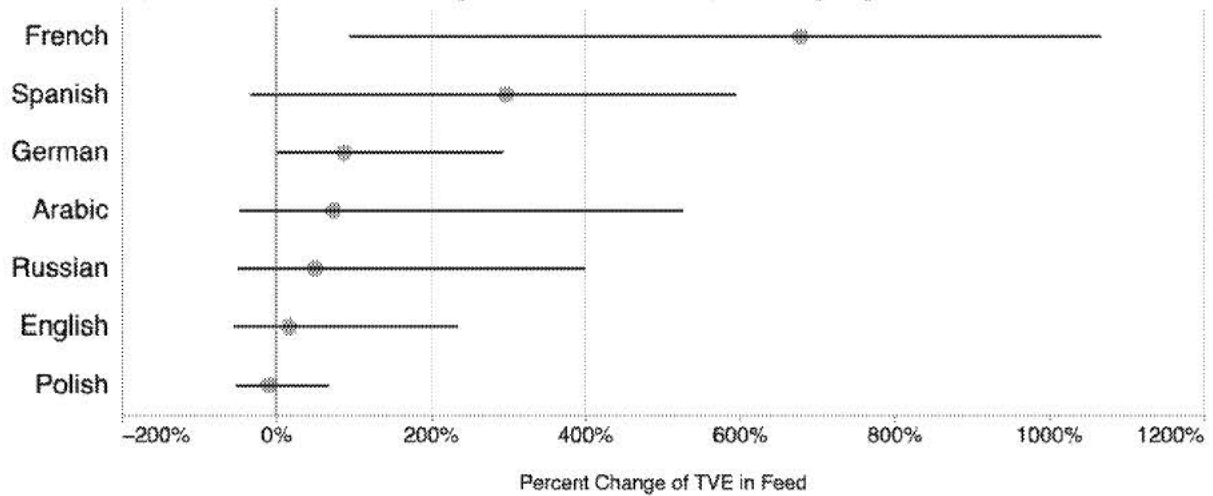


Figure A.6.11.7

This chart shows the percentage change for the amount of Bad (TVE) Content per language from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

Italian is missing from the above chart because there did not exist any Bad Content in feeds for the first search stage within the sample.

6.11.8 Amplification per Platform

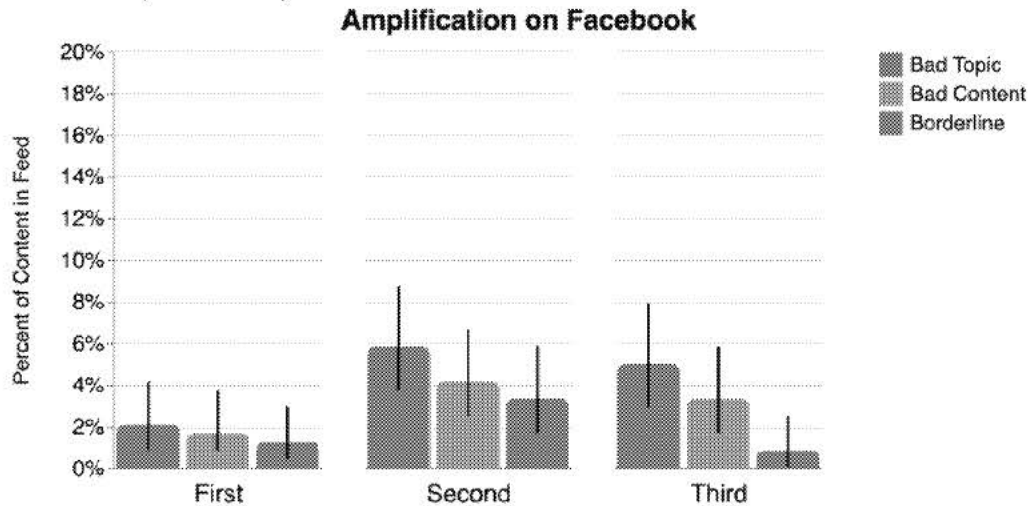


Figure A.6.11.8a

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Facebook for each stage. The black lines in the bars represent 90% confidence intervals.

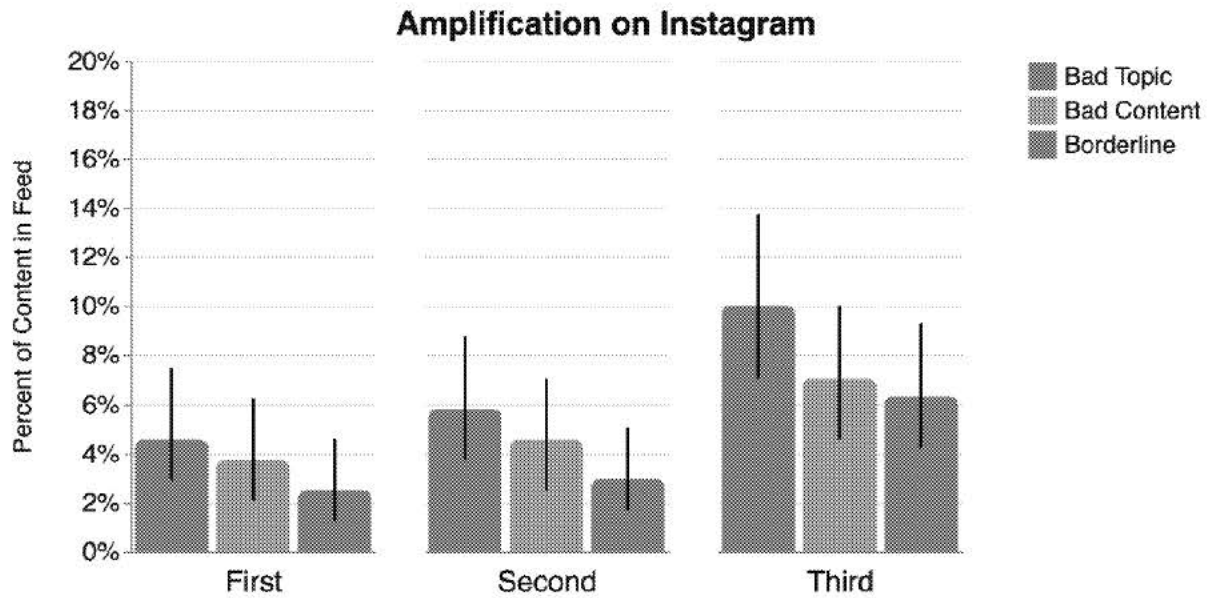


Figure A.6.11.8b

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Instagram for each stage. The black lines in the bars represent 90% confidence intervals.

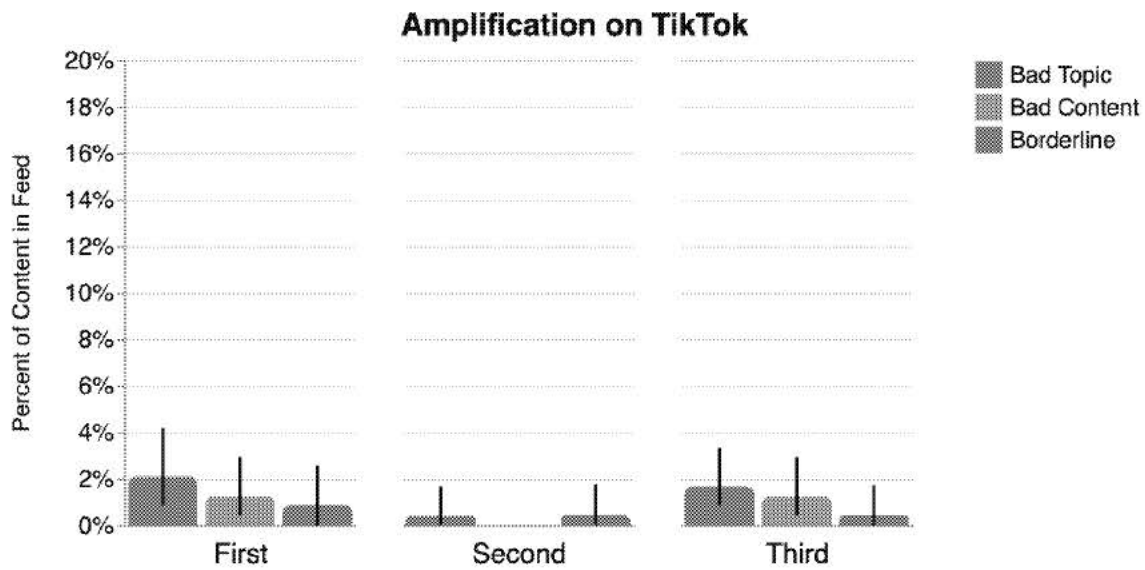


Figure A.6.11.8c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on TikTok for each stage. The black lines in the bars represent 90% confidence intervals.

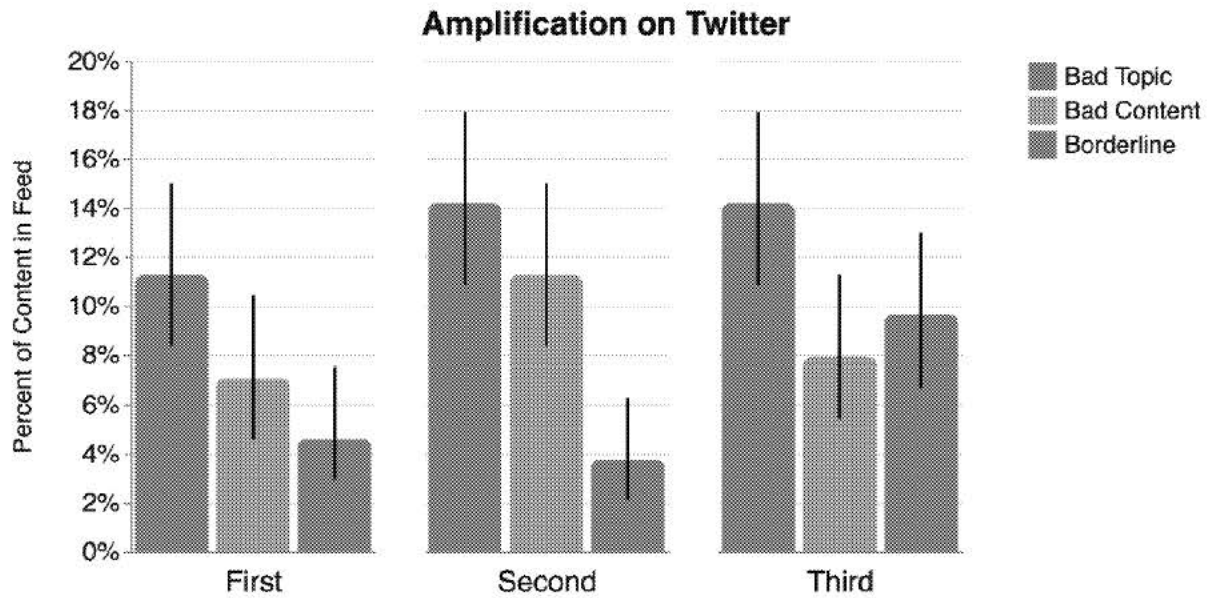


Figure A.6.11.8d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Twitter for each stage. The black lines in the bars represent 90% confidence intervals.

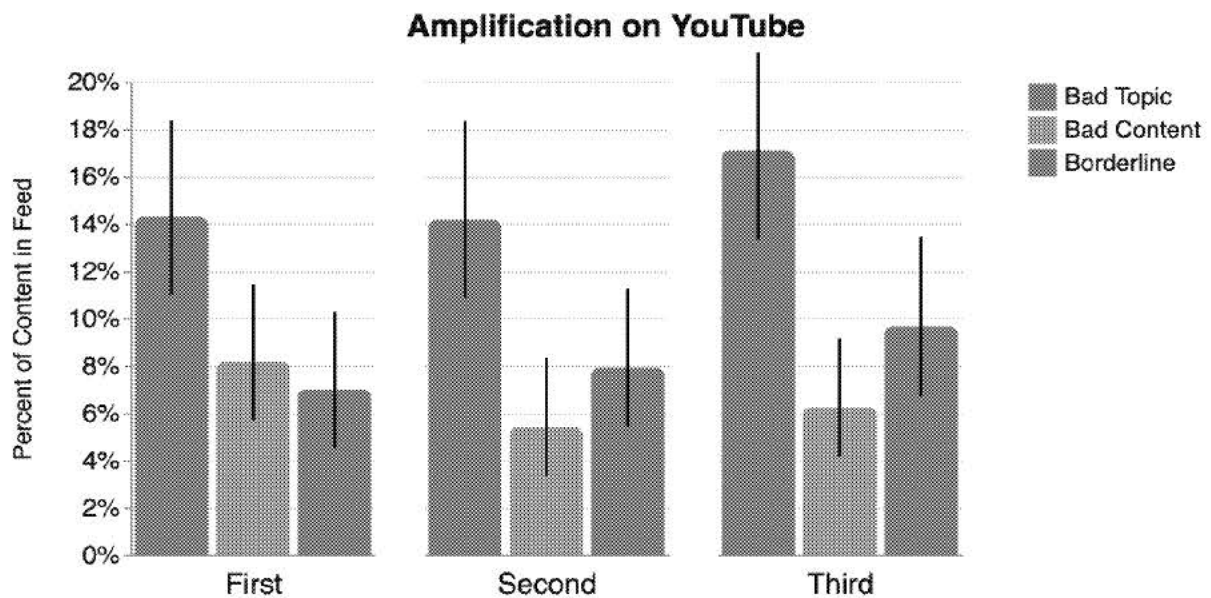


Figure A.6.11.8e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on YouTube for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.9 Amplification per Language

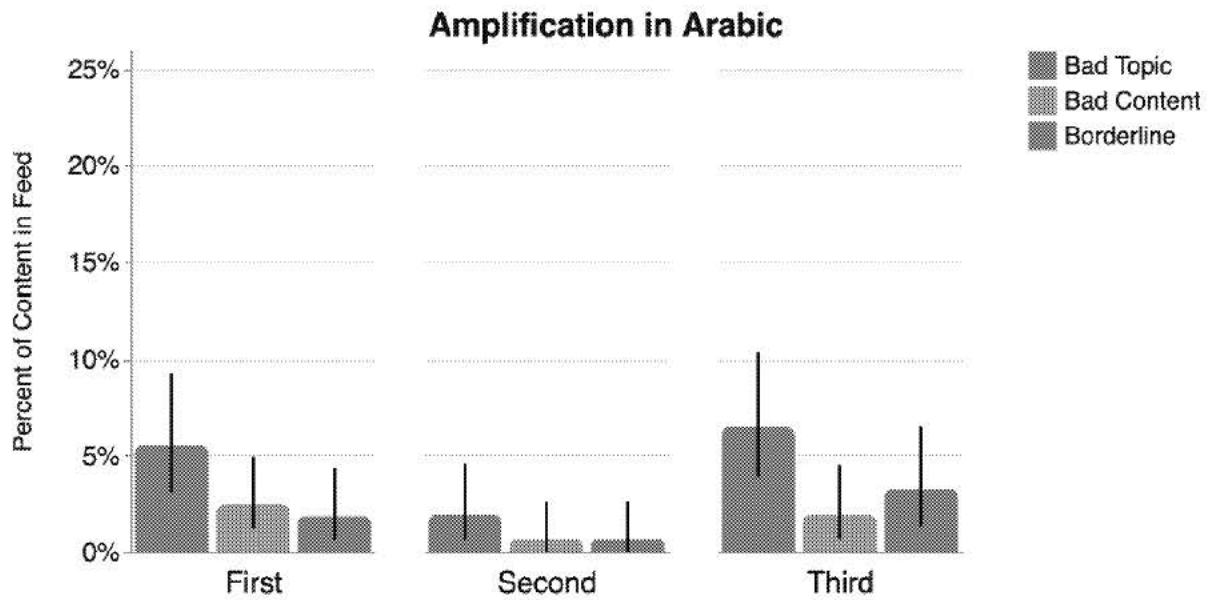


Figure A.6.11.9a

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Arabic for each stage. The black lines in the bars represent 90% confidence intervals.

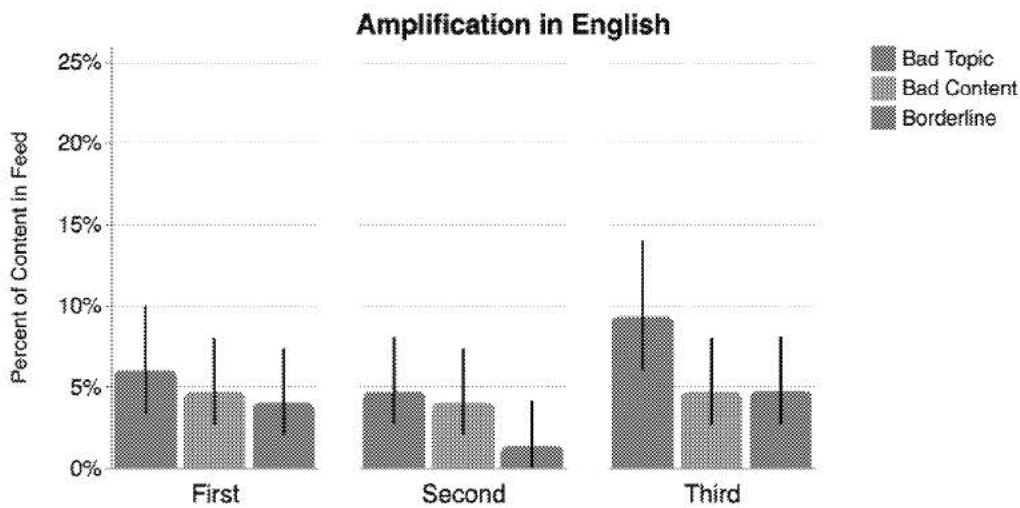


Figure A.6.11.9b

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in English for each stage. The black lines in the bars represent 90% confidence intervals.

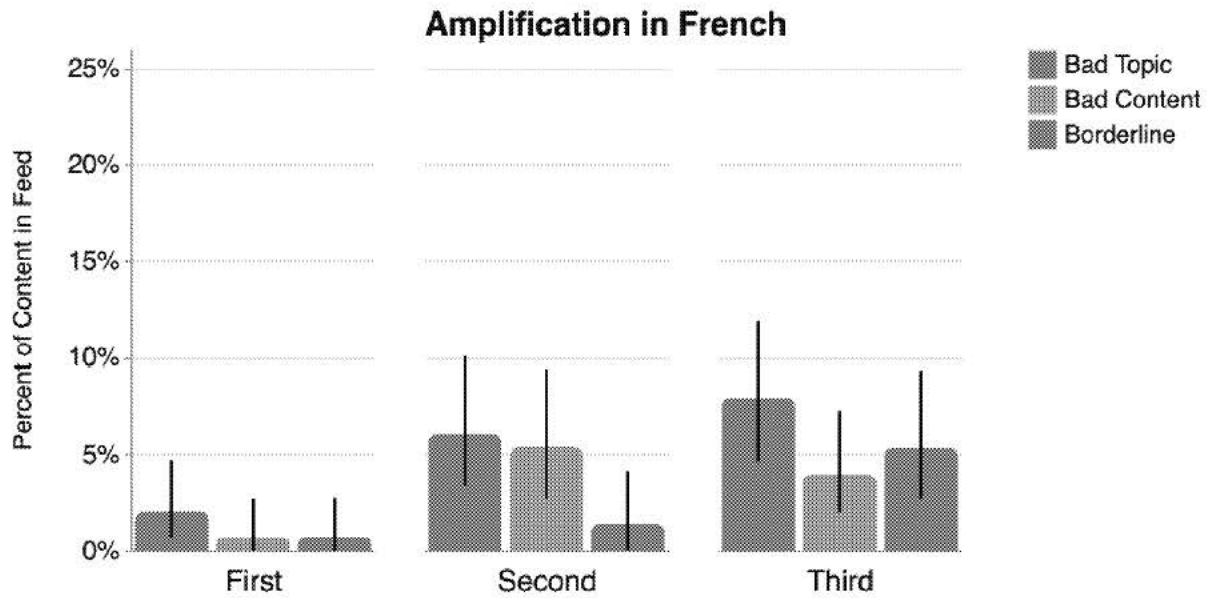


Figure A.6.11.9c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in French for each stage. The black lines in the bars represent 90% confidence intervals.

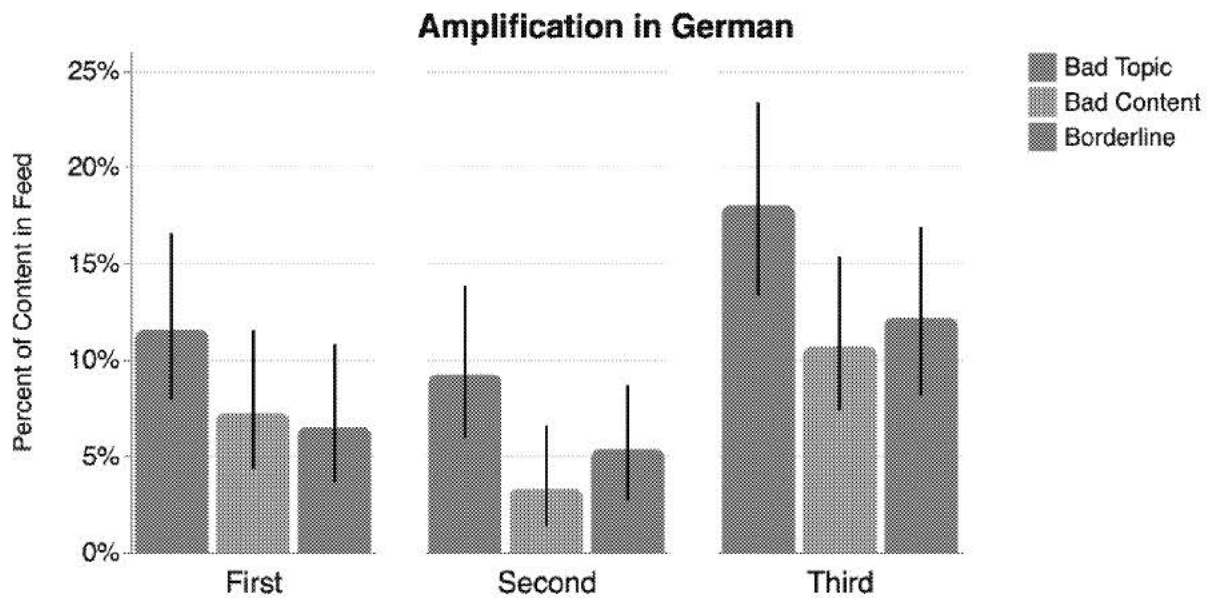


Figure A.6.11.9d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in German for each stage. The black lines in the bars represent 90% confidence intervals.

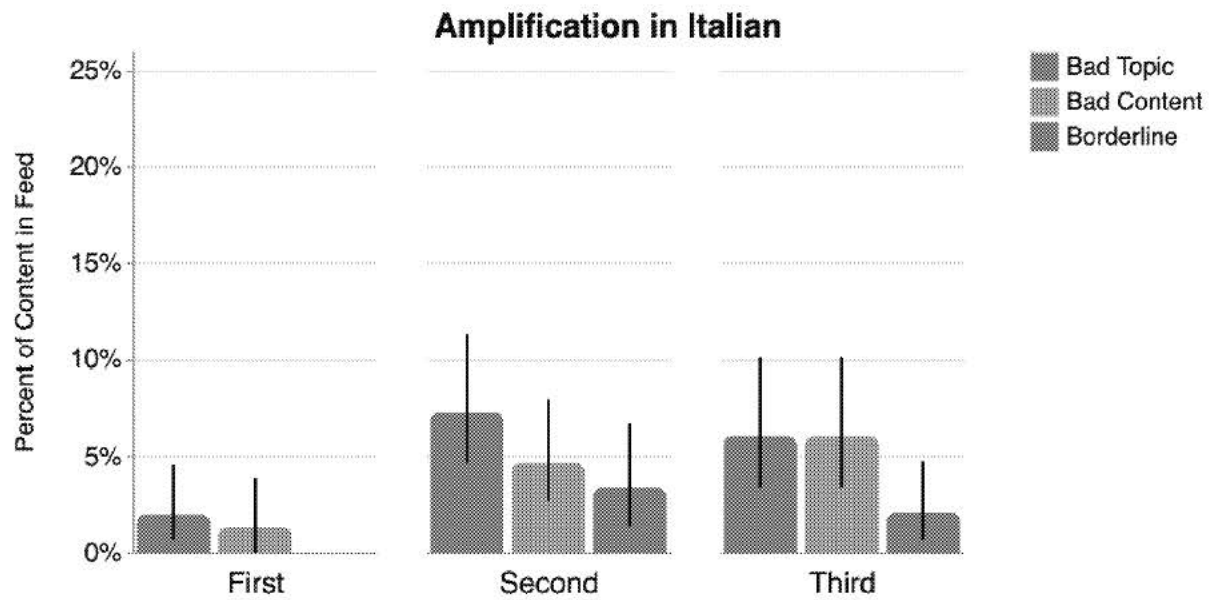


Figure A.6.11.9e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Italian for each stage. The black lines in the bars represent 90% confidence intervals.

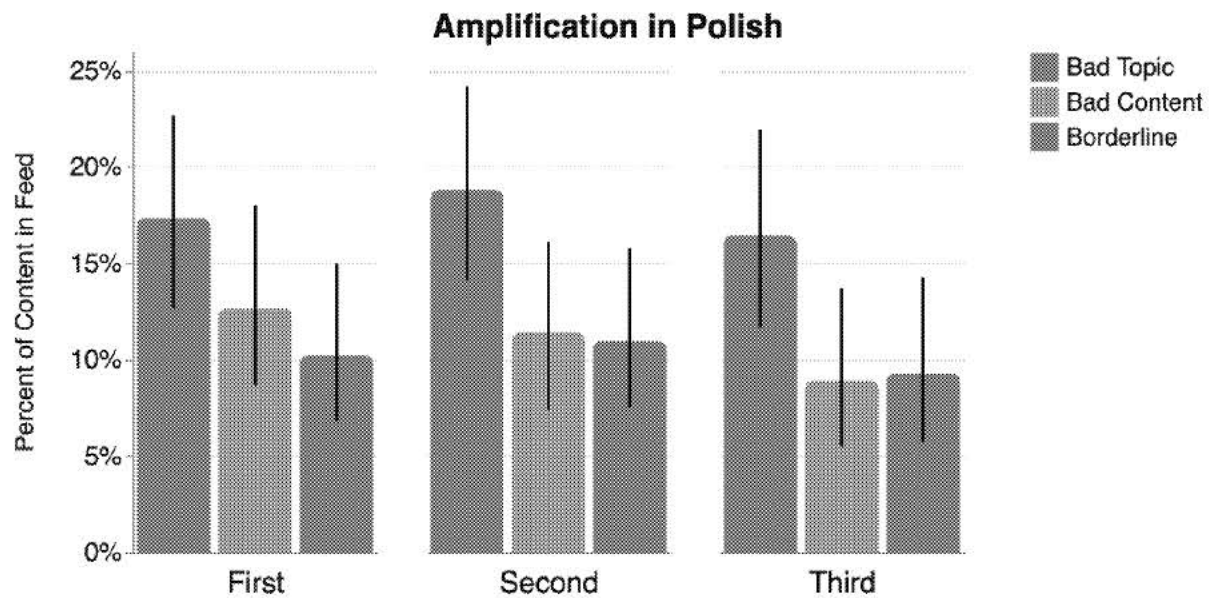


Figure A.6.11.9f

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Polish for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification in Russian

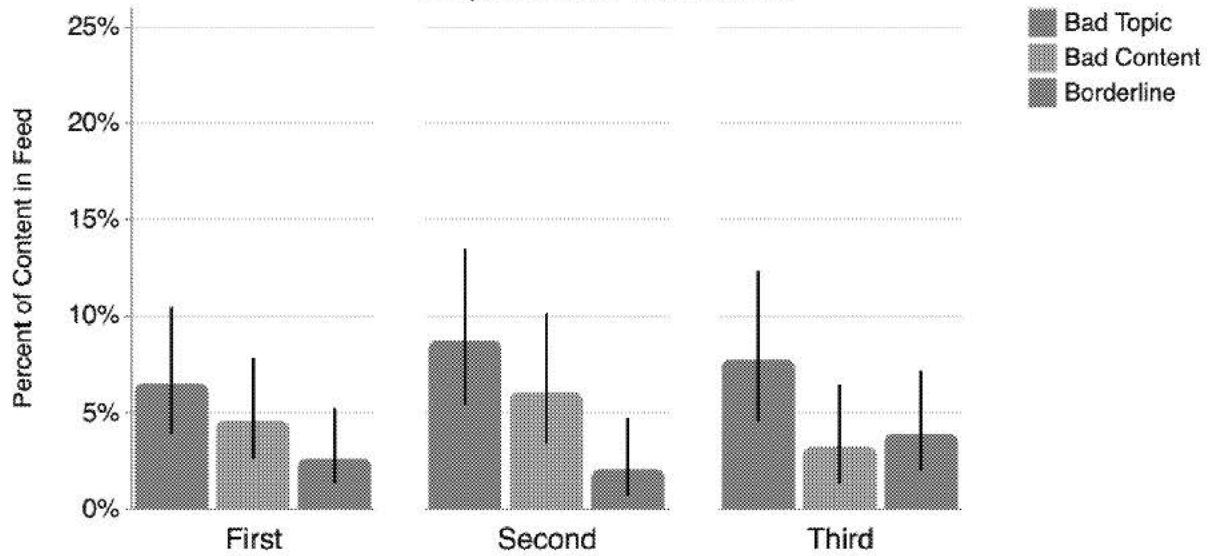


Figure A.6.11.9g

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Russian for each stage. The black lines in the bars represent 90% confidence intervals.

Amplification in Spanish

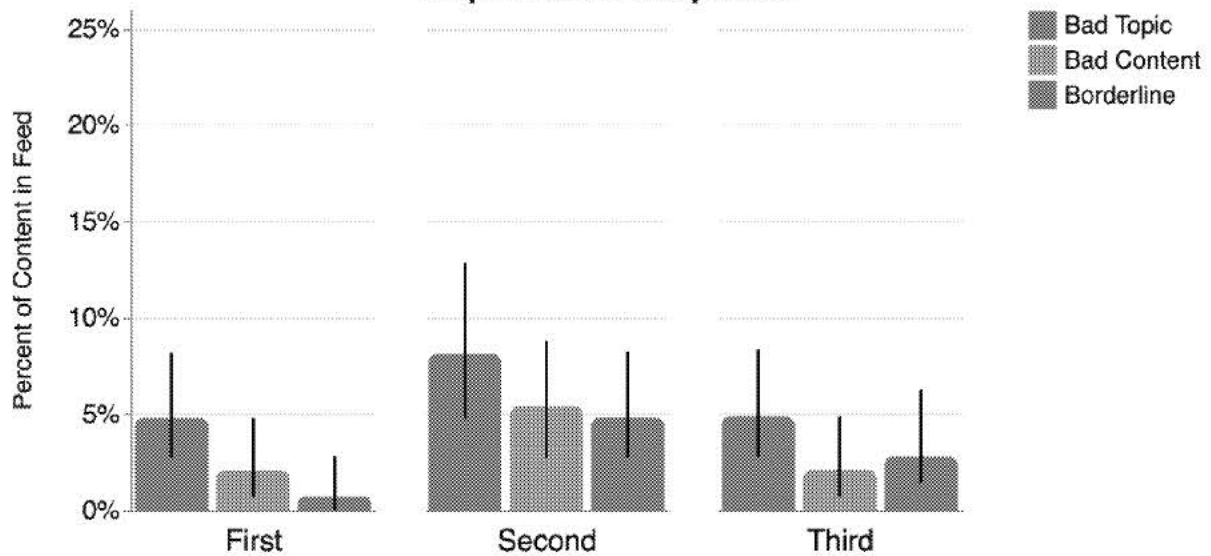


Figure A.6.11.9h

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Spanish for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.10 Amplification per TVE Type

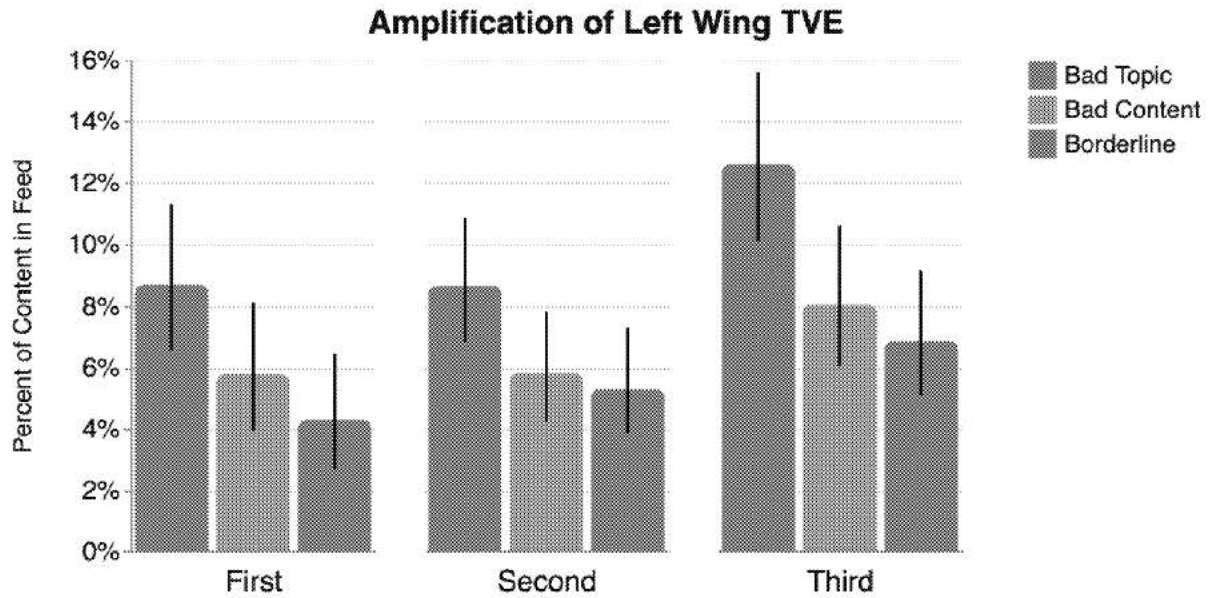


Figure A.6.11.10a

This graph illustrates the percentage of recommended Left-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

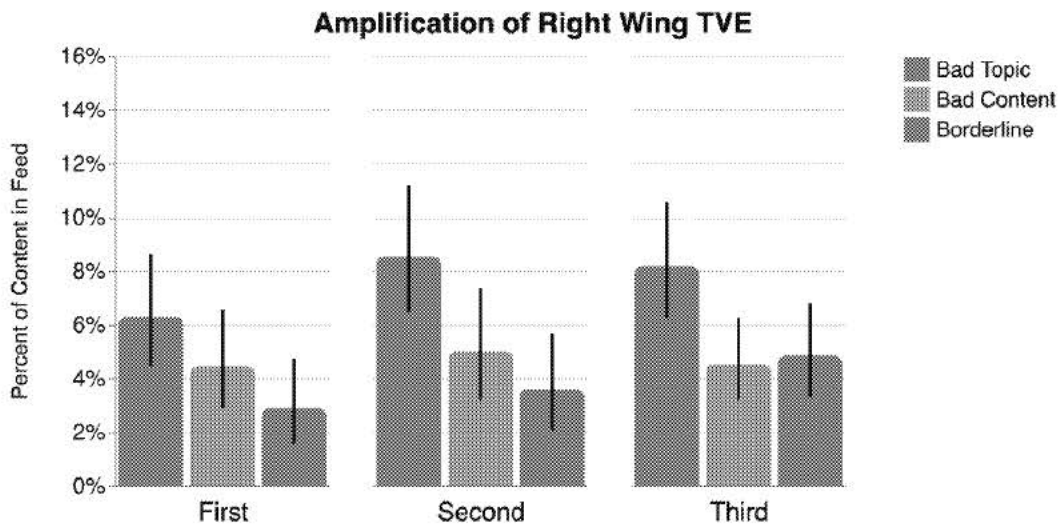


Figure A.6.11.10b

This graph illustrates the percentage of recommended National/Right-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

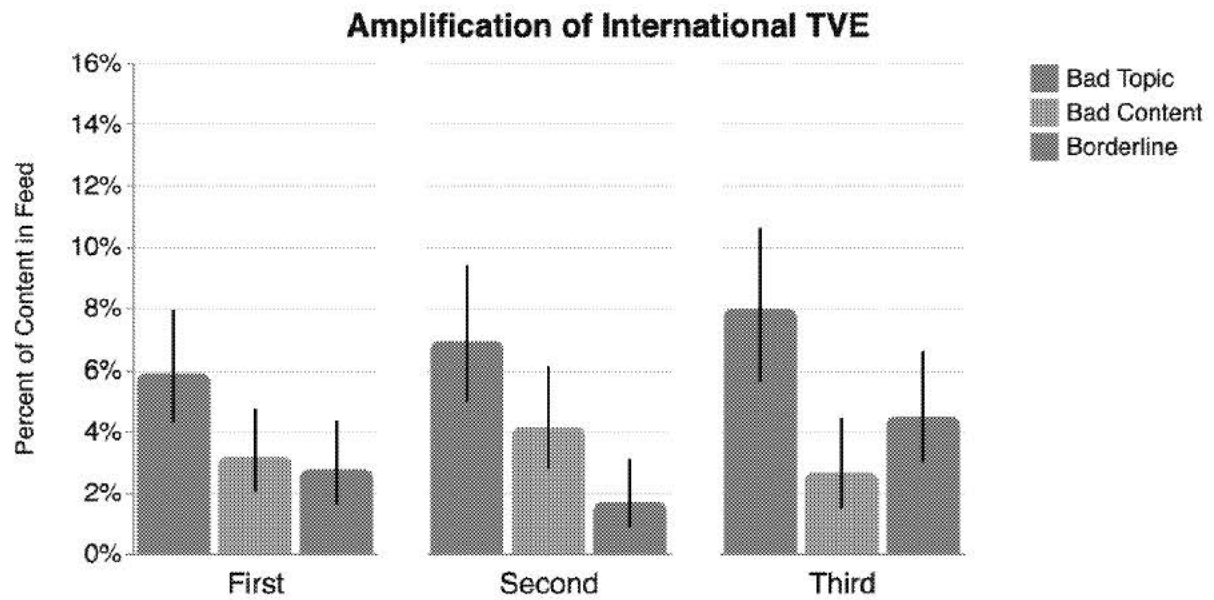


Figure A.6.11.10c

This graph illustrates the percentage of recommended international TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.11 Amplification P-values

Amplification Per Platform

Platform	Content Type	First/Second	First/Third	Second/Third
['Facebook']	Bad Topic	0.0352	0.0842	0.6875
['Facebook']	Bad Content	0.1041	0.2431	0.6317
['Facebook']	Borderline	0.1278	0.6538	0.0553
['Instagram']	Bad Topic	0.5387	0.0225	0.0913
['Instagram']	Bad Content	0.6486	0.1071	0.2435
['Instagram']	Borderline	0.7791	0.0447	0.0811
['TikTok']	Bad Topic	0.1007	0.7371	0.1782
['TikTok']	Bad Content	0.0826	1.0000	0.0826
['TikTok']	Borderline	0.5634	0.5634	1.0000
['Twitter']	Bad Topic	0.3384	0.3384	1.0000
['Twitter']	Bad Content	0.1142	0.7296	0.2156
['Twitter']	Borderline	0.6486	0.0328	0.0104
['YouTube']	Bad Topic	0.9701	0.3978	0.3799
['YouTube']	Bad Content	0.2306	0.4166	0.6976
['YouTube']	Borderline	0.6820	0.2908	0.5192

P-values on Amplification Average Percent by Content Type in Feed per Platform over time

Amplification per Language

108

Language	Content Type	First/Second	First/Third	Second/Third
['Arabic']	Bad Topic	0.0963	0.7269	0.0488
['Arabic']	Bad Content	0.1988	0.7540	0.3189
['Arabic']	Borderline	0.3441	0.4317	0.1014

Amplification per TVE Type

tve_type	Content Type	First/Second	First/Third	Second/Third
['International']	Bad Topic	0.5541	0.2520	0.5937
['International']	Bad Content	0.4608	0.6750	0.2793
['International']	Borderline	0.3139	0.1953	0.0317
['Left Wing']	Bad Topic	0.9888	0.0758	0.0538
['Left Wing']	Bad Content	0.9755	0.2102	0.1874
['Left Wing']	Borderline	0.4823	0.1127	0.3203
['National / Right']	Bad Topic	0.2485	0.2901	0.8635
['National / Right']	Bad Content	0.7282	0.9579	0.7547
['National / Right']	Borderline	0.6203	0.1643	0.3995

Table A6.11.11c

P-values on Amplification Average Percent by Content Type in Feed per TVE Type over time

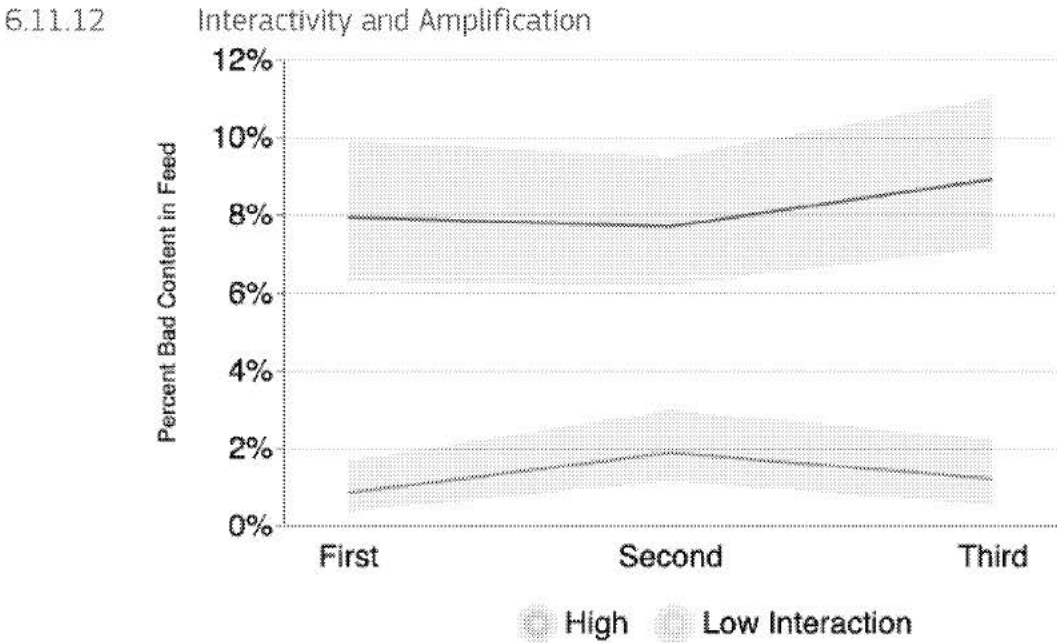


Figure A.6.11.12

This chart shows the average percentage of Bad Content for High and Low interaction over time. The shaded area represents the 90% confidence interval.

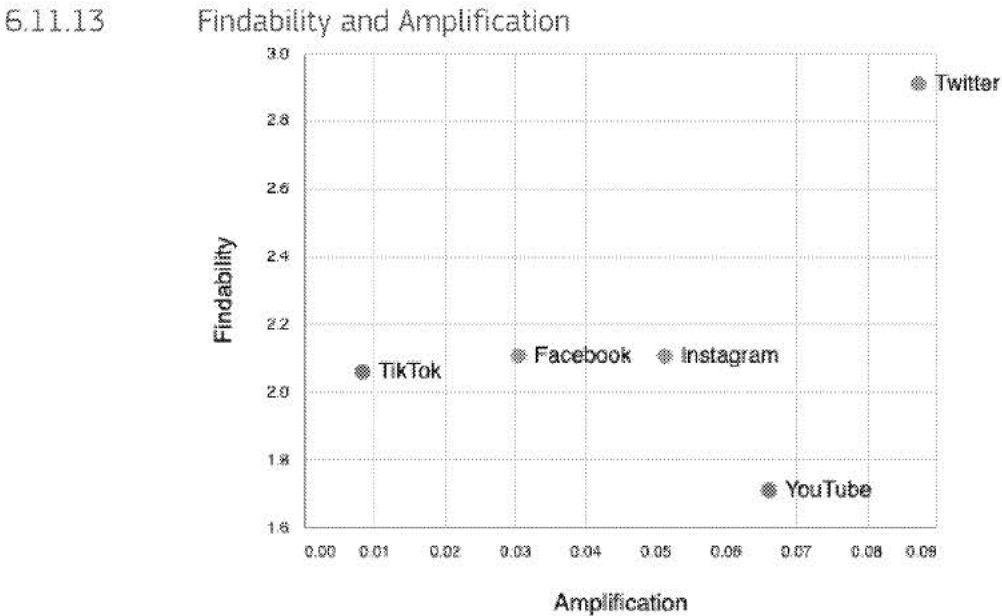


Figure A.6.11.13

This chart shows the comparison of the Findability score and the average percentage of Bad Content in the feed (here called Amplification). TikTok and YouTube are highlighted in red to contrast tech investment in feed vs search.

6.11.14 Engagement Ratio and Amplification

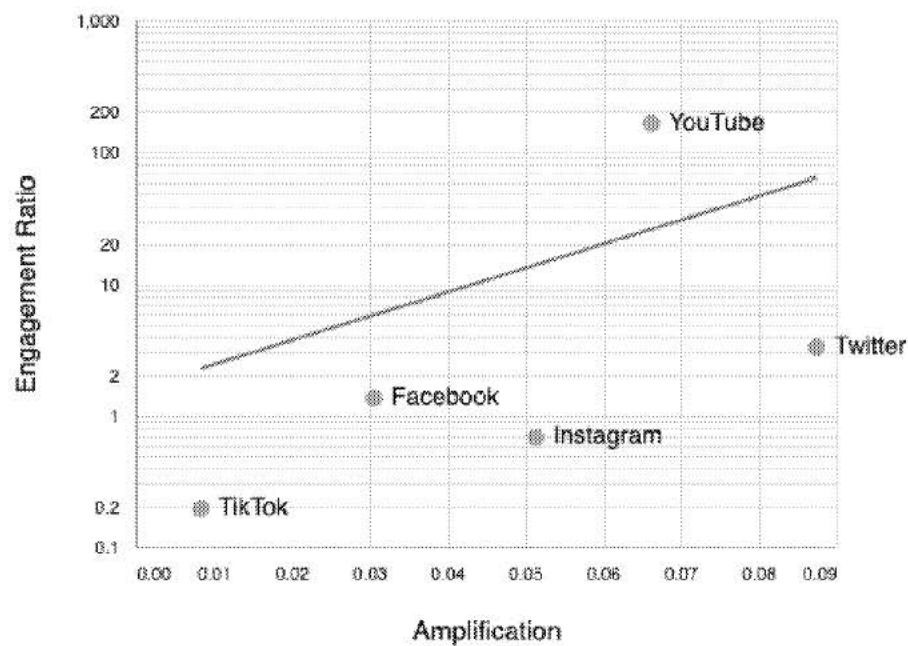


Figure A.6 11.14

This chart shows the positive correlation of the engagement and amplification. To normalise the engagement, an engagement ratio was used which is the median engagement (sum of the likes, shares, comments, and followers) of TVE content divided by the median engagement of non-TVE content. Amplification is the average proportion of the feed that contains TVE.

(note the log scale for the engagement ratio)

6.11.15 Removal Rate and Amplification

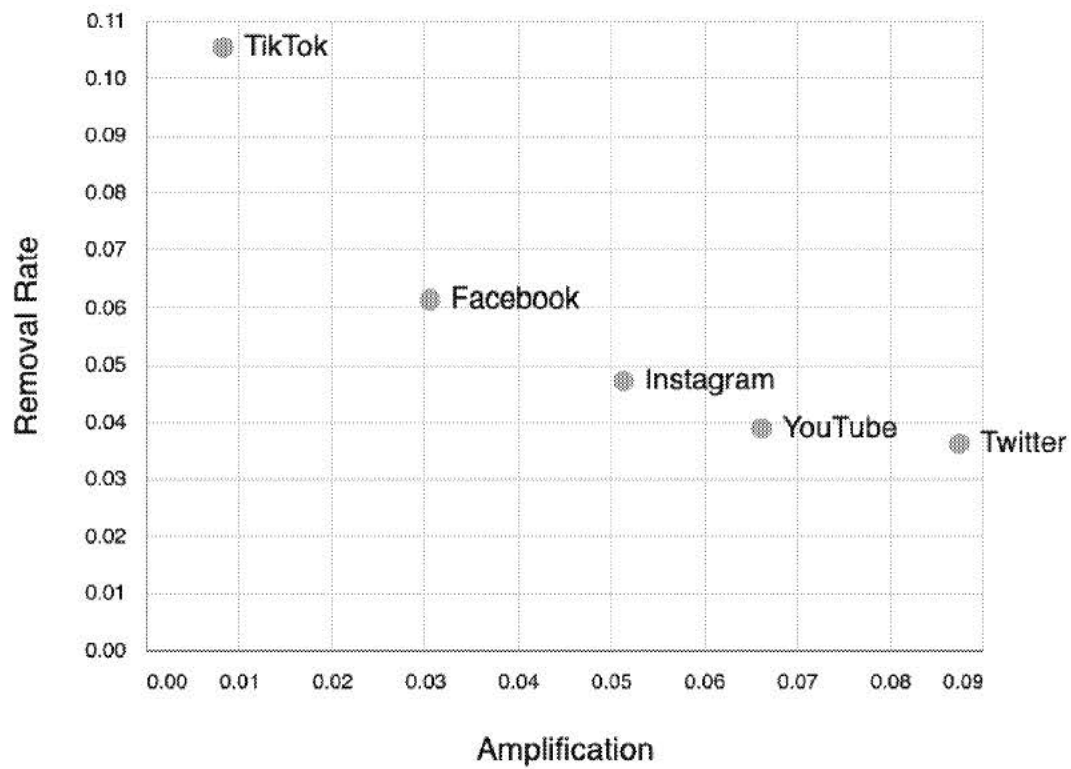


Figure A.6.11.15

This chart shows the inverse correlation of the removal rate and amplification.

6.12 Borderline Charts

6.12.1 Removal Rates per Platform

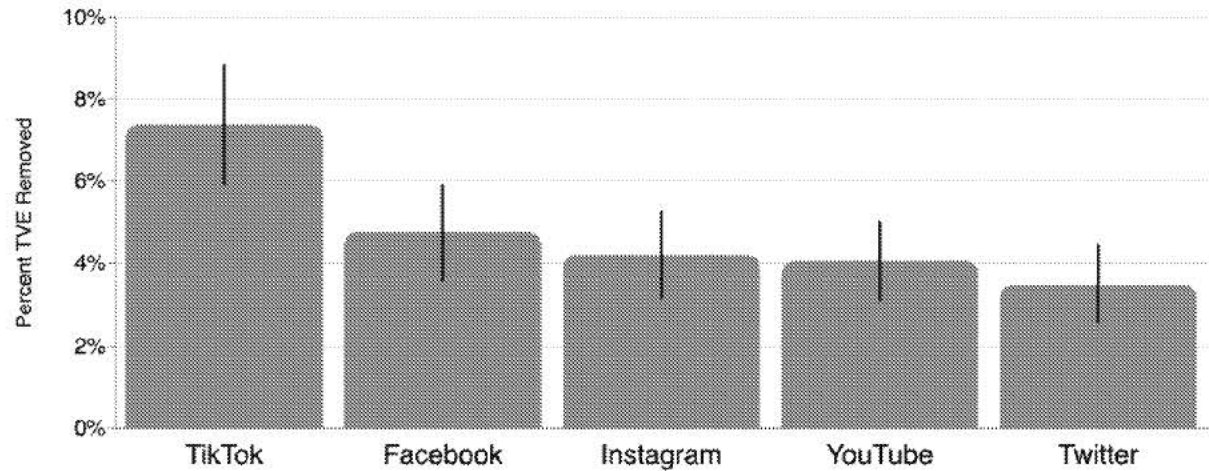


Figure A.6.12.1

This chart shows the percentage of Borderline TVE content that was removed by each platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.6397	0.9144	0.0505	0.6846
Twitter	0.4009	0.6201	0.0237	
TikTok	0.1604	0.0774		
Instagram	0.7283			

Table A.6.12.1

P-values on Removal Rates per Platform (Borderline)

$\alpha = 0.01$ (Bonferroni correction from 0.05)

6.12.2 Removal Rate per Language

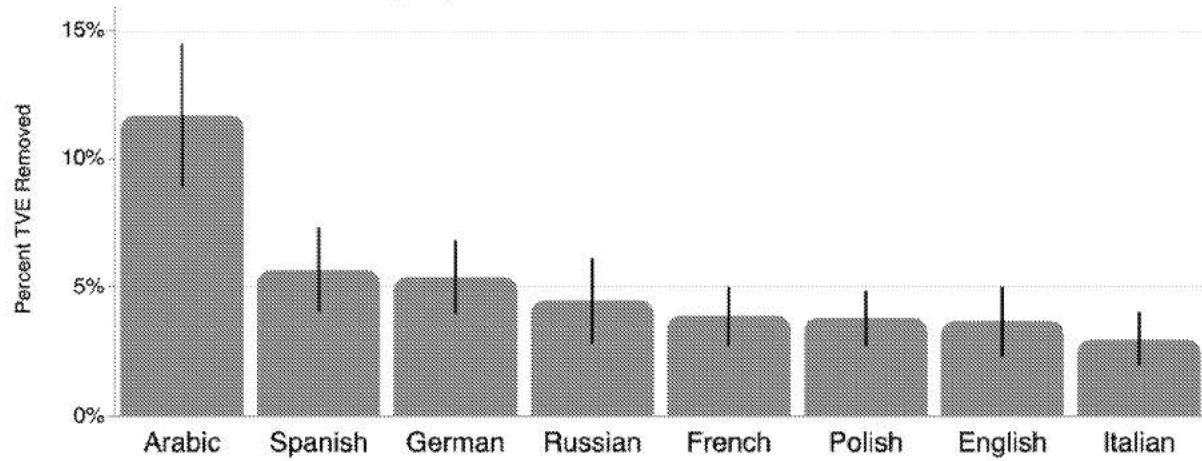


Figure A.6.12.2

This chart shows the percentage of Borderline TVE content that was removed for each language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0507	0.3527	0.3581	0.8964	0.1516	0.3207	0.6160
Russian	0.0226	0.7029	0.7601	0.6869	0.4226	0.7209	
Polish	0.0014	0.9431	0.9576	0.3679	0.5937		
Italian	0.0005	0.6862	0.5672	0.1741			
German	0.0281	0.3993	0.4095				
French	0.0023	0.9076					
English	0.0052						

Table A.6.12.2

P-values on Removal Rates per Language (Borderline)

$\alpha = 0.00625$ (Bonferroni correction from 0.05)

6.12.3 Removal Rate per TVE Type

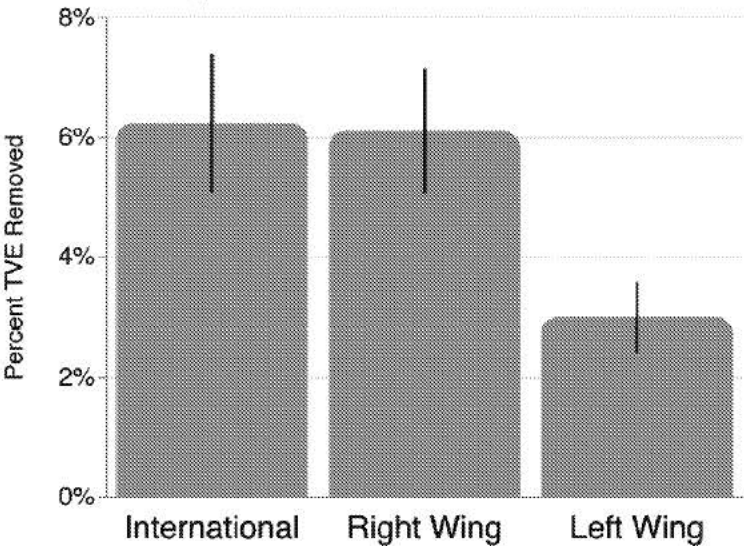


Figure A.6.12.3

This chart shows the percentage of Borderline TVE content that was removed for each TVE Type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.9357	0.0051
Left Wing	0.0056	

Table A.6.12.3

P-values on Removal Rates per TVE Type (Borderline)

alpha = 0.017 (Bonferroni correction from 0.05)

6.12.4 Removal Time per Platform

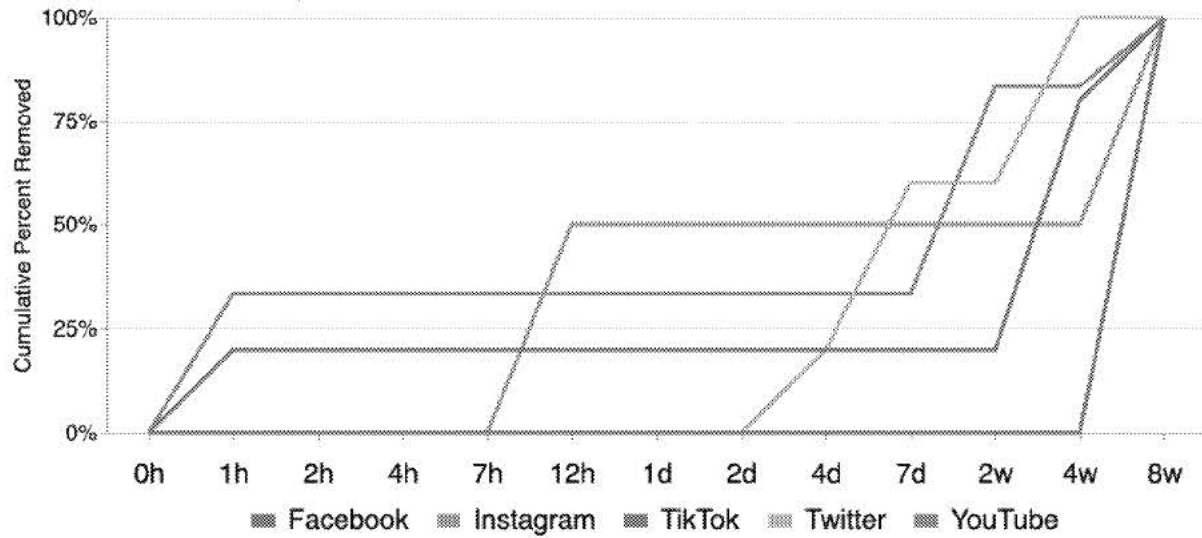


Figure A.6.12.4

This chart shows how long it took for each platform to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.5 Removal Time per Language

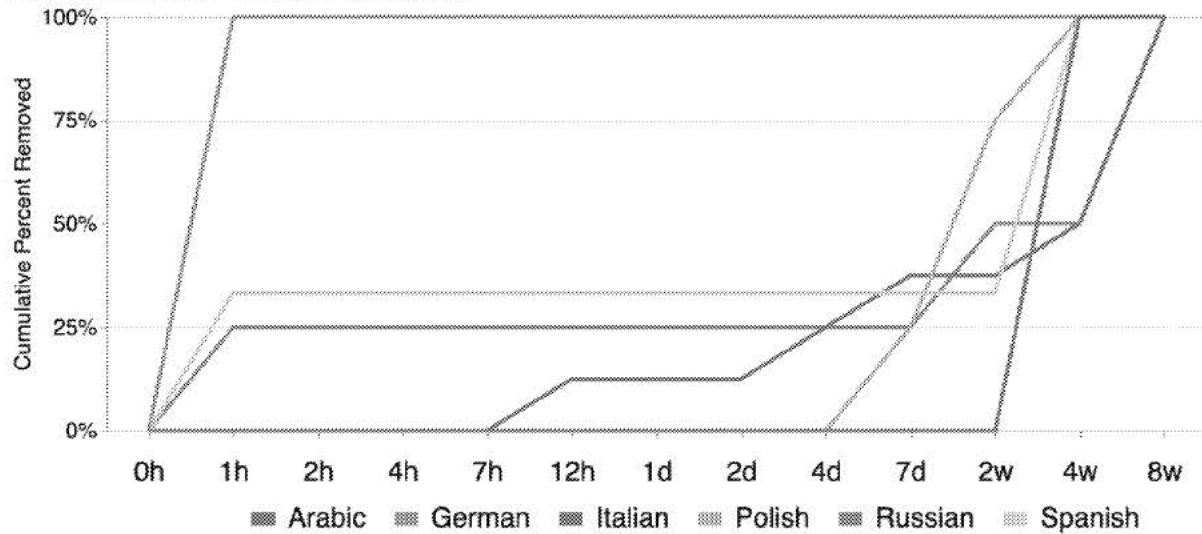


Figure A.6.12.5

This chart shows how long it took for each language to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.6 Removal Time per TVE Type

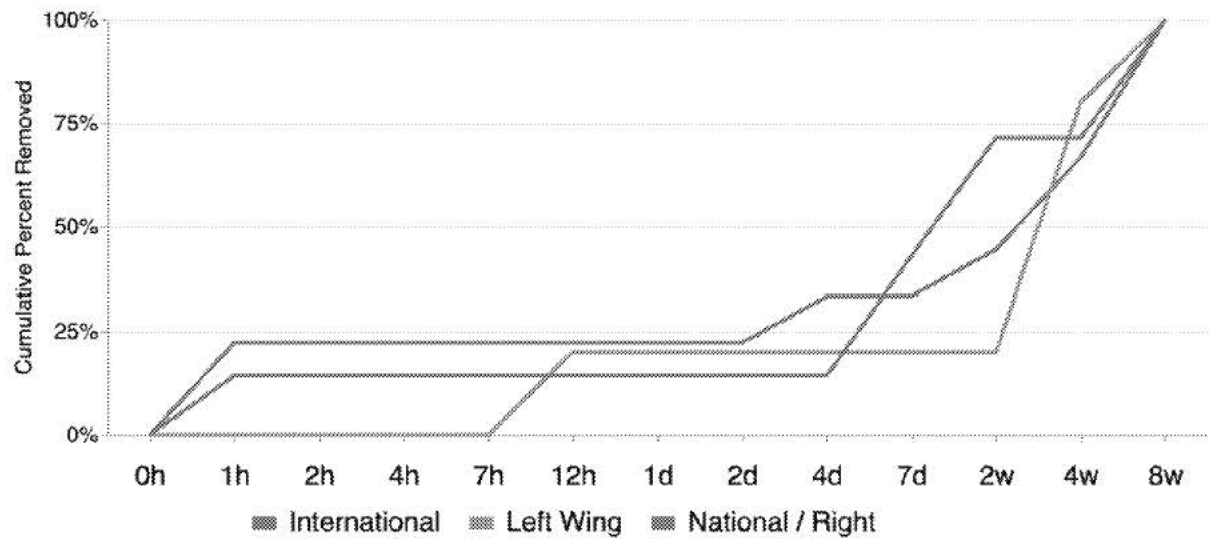


Figure A.6.12.6

This chart shows how long it took for each TVE Type to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.13 Task 3: List of References

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410-428.

Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78, 413-418.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*, 10(21), 7748.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945.

Grimmelmann, J. (2015) The virtues of moderation. *Yale Journal of Law & Technology* 17: 42.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3), 261-273.

Mohamed, M. H., Khafagy, M. H., & Ibrahim, M. H. (February 2019). Recommender systems challenges and solutions survey. In 2019 international conference on innovative trends in computer engineering (ITCE) (pp. 149-155). IEEE.

Murthy, D. (2021). Evaluating platform accountability: terrorist content on YouTube. *American behavioral scientist*, 65(6), 800-824.

Suhaim, A. B., & Berri, J. (2021). Context-aware recommender systems for social networks: review, challenges and opportunities. *IEEE Access*, 9, 57440-57463.

Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7, 439-457.

END OF THE DOCUMENT

Exhibit 41

EUIF Questionnaire - Algorithmic Amplification_TT response

Questionnaire on Activities related to Algorithmic Amplification

This questionnaire follows the first EUIF informal focus group meeting to further address algorithmic amplification.

This questionnaire shall serve as a basis to establish a comprehensive set of good practices of policies implemented and measures taken by companies to address unintentional spread of borderline, violent extremist and terrorist content as well as to prevent targeted exploitation by malicious actors to disseminate said content. This includes measures and policies in place such as options to opt-out of automated recommendation of content, excluding users who are violating terms of services from recommender systems or limiting the repetition of videos with the same content.

We kindly ask you to explain the purposes of the measures taken/policies implemented in the EU and where available evaluation of effectiveness and impact of those. We kindly ask you to name and explain the policy/measures in this survey and refrain from simply providing a link to the general policy/measure.

This work builds on the EUIF's meeting on algorithmic amplification (September 2021, September 2022) and the EUIF Study on role and effects of the use of algorithmic amplification (2023). It responds to activity 1.C4 of the EU Internet Forum Activities 2024.

Company:

If you cover several platforms, please specify which measures/policies apply to which platform.

TikTok's responses in yellow.

- 1. What policies and/or measures have you implemented since July 2022 to address (1) unintentional spread and (2) exploitation of algorithmic recommender systems by malicious actors for terrorist content on your platform?**

Please specify if this relates to recommender algorithms, search algorithms or content moderation algorithms.

2. What policies and/or measures have you implemented since July 2022 to address 1) unintentional spread and (2) exploitation of algorithmic recommender systems by malicious actors for violent extremist content on your platform?

Please specify if this relates to recommender algorithms, search algorithms or content moderation algorithms.

a. How do you define violent extremist content

Answer Relating to Questions 1 and 2:

Terrorist content and violent extremism are categories of illegal content, and as such, are prohibited on TikTok. The term "violent extremist" is also defined in our Community Guidelines (under Violent and Hateful Organisations and Individuals). TikTok maintains a zero-tolerance policy on violent extremism. Individuals or groups that threaten violence or attempt to incite violence risk the safety of our community. Moreover, our Community Guidelines clearly outline that we do not allow people to use our platform to threaten or incite violence, or to promote violent extremist organizations, individuals, or acts.

We use a combination of technology, safety and security professionals, and threat detection partners to help us enforce our policies. An account that attempts to promote or glorify off-platform violence or threaten public safety can be banned. When appropriate, we will report threats to relevant legal authorities. We may consider off-platform behaviour to identify violent extremist organisations and individuals and take action on their accounts.

We regularly provide more information on our enforcement efforts via our quarterly Community Guidelines Enforcement Report.

In addition, and as required by Article 16 of the Digital Services Act, TikTok has put in place a specific reporting channel for illegal content in Europe, under which users can report content relating to violent or organised crime, national security-related offences, and terrorist content and offences. In addition, designated Trusted Flaggers may submit reports using the illegal content reporting mechanism on TikTok in line with Article 22 of the DSA, and Member State authorities may use either the TikTok designated single point of contact or a designated form to report illegal content to TikTok, in line with Article 9 of the DSA. TikTok publishes a bi-annual DSA Transparency Report which provides information on the company's enforcement actions taken with respect to illegal content in Europe.

To further strengthen our commitment to user safety and human rights, in September 2022 TikTok announced its membership of Tech Against Terrorism, which brings together technology companies, civil society, and academics over the shared goal of countering violent extremism. By joining Tech Against Terrorism, TikTok earned their Trustmark, which is proof that we adhere to membership requirements, which include commitments to exploring new technical solutions and working with civil society on combating violent extremism. The

Trustmark also demonstrates our strict policies and enforcement against violent extremism and our commitments to human rights and transparency.

Through our membership, we're eager to continue to learn from others as we strive to meet and exceed best practices in tackling this issue. It also provides us with access to training and detection resources to help us effectively respond to such content and actors.

3. What policies and/or measures have you implemented since July 2022 to address 1) unintentional spread and (2) exploitation of algorithmic recommender systems by malicious actors for borderline content on your platform?

a. How you define borderline content?

Please specify if this relates to recommender algorithms, search algorithms or content moderation algorithms.

We do not use the term "borderline content" in our Community Guidelines.

In general, TikTok has put in place a variety of policies and measures to prevent malicious actors from exploiting our algorithm, and to provide our user community with tools to customise their For You Feed viewing preferences and ensure that they have an enjoyable experience on the service.

For instance, certain categories of content may be ineligible for recommendation, and viewers can use our "not interested" feature to automatically skip videos they want to see less of.

To further empower users with ways to customise their viewing experience, in summer 2022 we rolled out a tool that people can use to automatically filter out videos with words or hashtags they don't want to see in their For You or Following feeds (keyword filtering).

Around the same time, TikTok also introduced an option for users to refresh their For You Feed if they feel that their recommendations are no longer relevant or provide enough variety.

Finally, as required by the DSA, TikTok also offers users in Europe a new way to discover content on TikTok by allowing them to turn off personalisation. When enabled, this means their For You and LIVE feeds will instead show popular videos rather than recommending content to them based on their personal interests.

4. Which commitments do you follow in relation to algorithmic amplification, under the

a. Code of Conduct on Hate Speech?

b. The Code of Practice on Disinformation?

c. Any other initiative (please provide more information on that initiative)?

As both the CoC on Hate Speech and the COPD on Disinfo are currently undergoing revision and will be converted into DSA Codes as of early 2025 (tbc), we recommend

discussing this topic during another EUIF meeting as needed, when the new frameworks are in place.

5. Do you have any policies/measures in place to pro-actively expose users to a diversity of content (to prevent the creation of echo chambers)?

To address repetitive patterns, in early 2023 TikTok rolled out new tools and policies to diversify recommendations, to ensure that we avoid recommending a series of similar content on topics that may be fine as a single video but potentially problematic if viewed repeatedly. TikTok also does not recommend videos from the same creator or that uses the same sound/song in sequence.

6. Do you have any policies/measures in place related to your algorithmic recommender system for crisis situation, such as terrorist attacks with an online dimension or violent extremist protests?

7. What you do in case of crisis, such as the Southport protests in the UK (August 2024)? Do you have any temporal measures, such as de-ranking material for crisis?

8. Are there any situations in which you do not remove terrorist content but take other measures, such as de-ranking, to moderate the content? If so, please explain why?

No, terrorist content that is identified as such, is removed.

9. Which human-rights measures are you taking when deranking content and how do you ensure transparency of your measures?

It is not clear what is meant by the term "human-rights measures", or the kinds of categories of content this question is referring to.

Regarding the question of transparency, users are provided with a notification where we determine that they have violated our Policies and applicable laws, and they are provided with an opportunity to appeal against any action we have taken. We provide information to users about how to appeal a decision in relation to a report they have made, or how to appeal a decision affecting their content or account [here](#). We report on the number of appeals, and the action we take in response to those appeals, in our DSA Transparency Reports. In addition, TikTok has a feature known as Account Check, which allows users to check their account status in-app, to check if content has been removed for violating our policies and applicable laws. In this way, even if a user has missed a notification in their inbox relating to a content moderation decision, they can still check the status of their

account and its associated videos in-app.

10. Do you collaborate with research/academia/ civil society organisation to enable a better understanding of the possible effects of algorithmic recommender systems on a user's journey of radicalisation? If so, how?

11. Do you collaborate with research/academia/ civil society organisation on online prevention efforts, such as via re-direct campaigns, counter-narrative campaigns, amplification of authoritative sources during a crisis?

Regarding questions 10 and 11, we collaborate on an ongoing basis with researchers, academia, and civil society, across a variety of different initiatives, and to better develop our own approaches to violative content.

For example, TikTok developed an ongoing partnership with Tech Against Terrorism which ran throughout most of 2024, through which we held monthly workshops to develop our understanding of terrorist and violent extremist behaviour and tactics online, and develop awareness of adversarial shifts and mitigation strategies.

Regarding crises, we continuously consult our Safety Advisory Council experts in Europe, the Middle East and Asia Pacific, particularly during critical moments and crises such as the Israel/Palestine and Ukraine/Russia wars.

In terms of regular consultation with academia and civil society in the violent extremism space, we have engaged with stakeholders such as the International Committee of the Red Cross (Israel/Palestine), the Centre for Humanitarian Dialogue (VBDA policies in times of conflict), the VOX-Pol Institute, GlitterPill (Conflict Zones and Russia/Ukraine), the Global Internet Forum to Counter-Terrorism (On TikTok Live and high harms), and GNET (on Ukraine).

More broadly, we provide data access via our Research Tools and have also presented and demonstrated the tool at various conferences, including Terrorism and Social Media in Swansea (19 June 2024).

12. Do you utilise any of the EUIF deliveries, such as the Handbook on Borderline Content or the Knowledge Package on violent-right wing extremist, groups, symbols and manifestos for your automates systems? Please provide details, and reasons if not.

The handbook has been shared internally with our teams, to better inform the enforcement of our policies.

13. **Any addition information that you would like to share on policies/measures that you have taken to address the spread of borderline, violent extremist and terrorist content on your platform?**

We thank you for your collaboration and taking the time to fill in the survey.

Section IV:

Hate Speech Code Exhibits

Exhibit 42



The ninth EU High level group on combating racism, xenophobia and other forms of intolerance

Online meeting, 7 July 2021

Presentations

DG JUST Unit Fundamental rights policy

Subgroup on combating illegal hate speech online

The Digital Service Act proposal

Redacted

European Commission (CNECT)



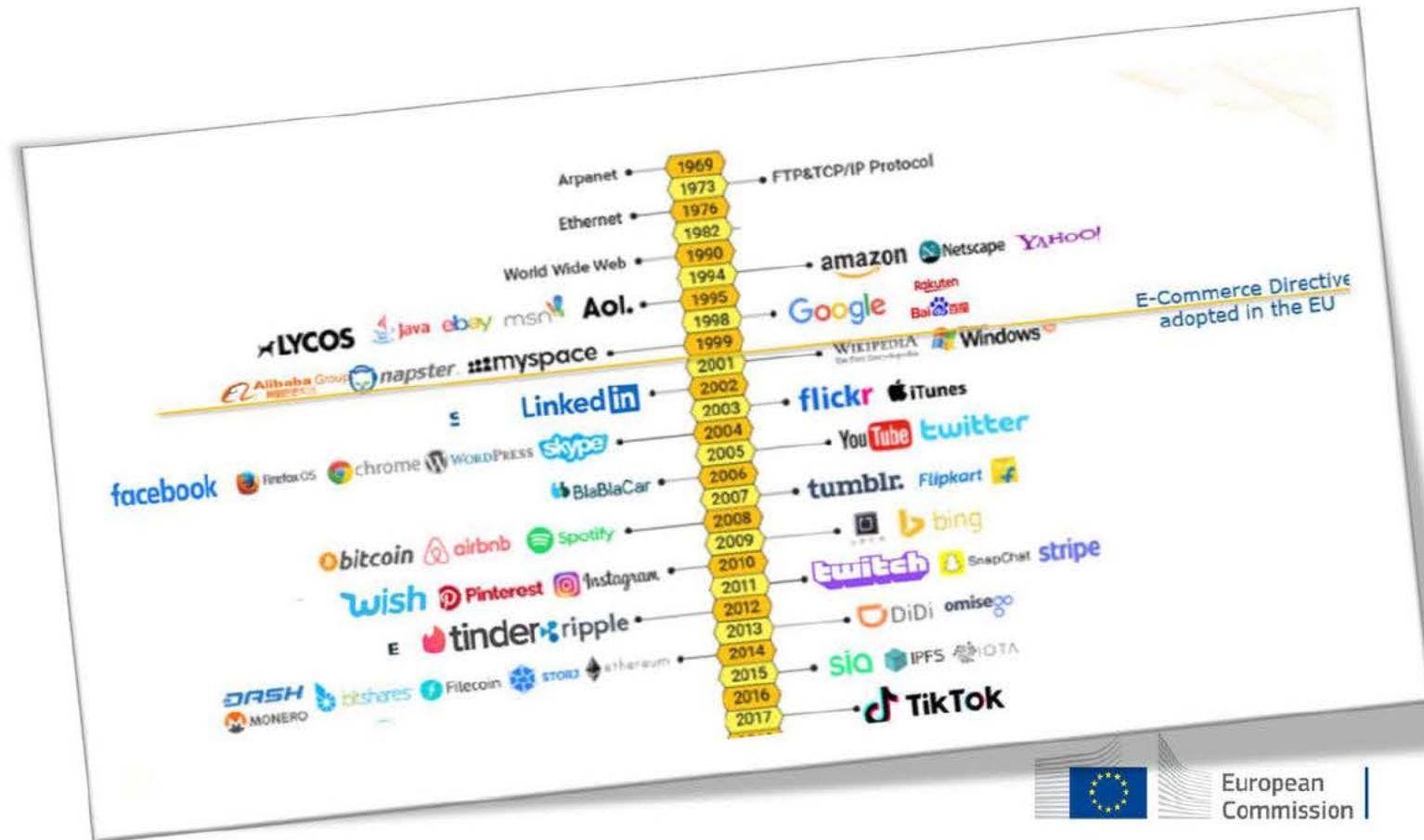
DIGITAL SERVICES ACT

Evolution and role of intermediary services

- A step-change in the nature, scale and role that digital services play in our economy and society.

- E-Commerce for example:

- 9% of Europeans shopped online in 2002
- 70% shop online today – and 40% of businesses that sell online, do so through online platforms



European
Commission

unsafe products

terrorist offences

copyright infringements

scams

hate speech

illegal apartments

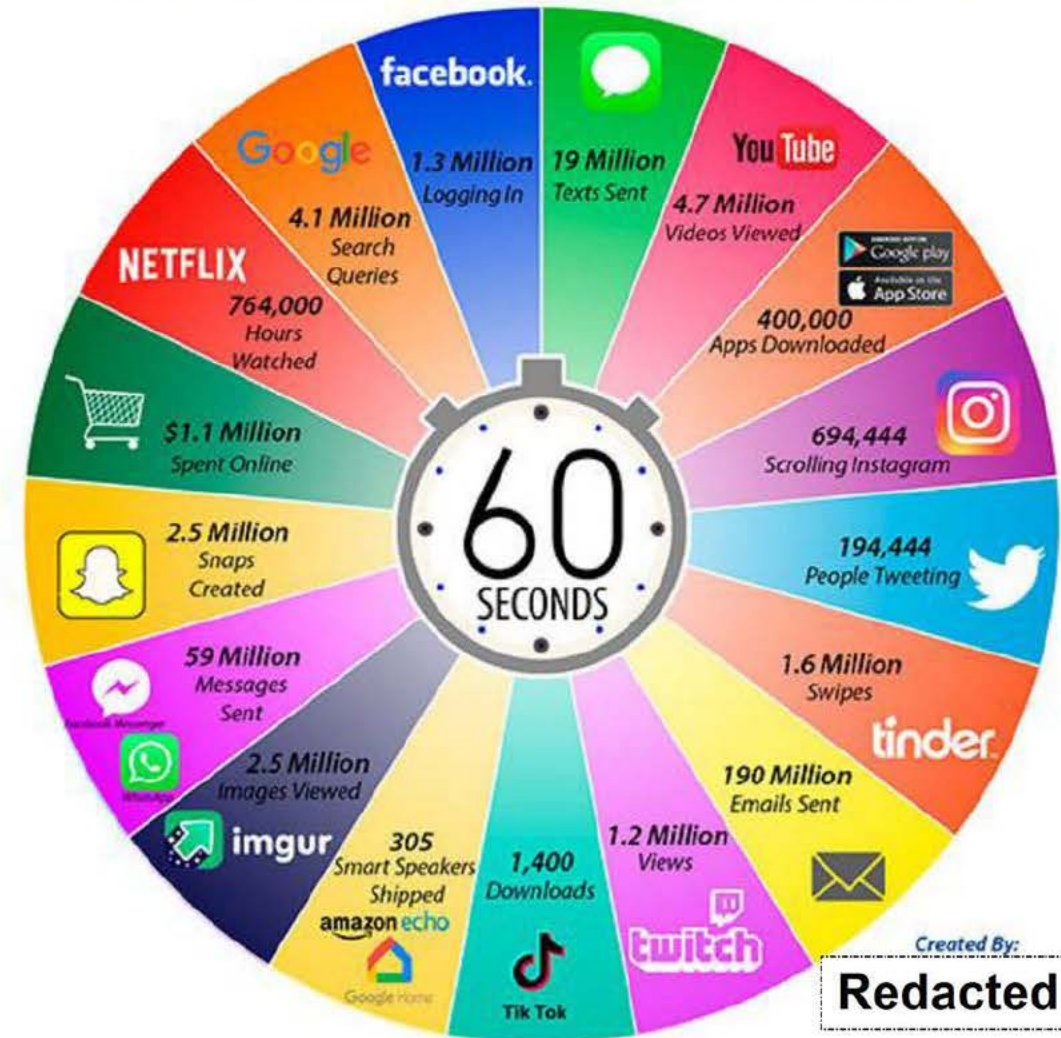
counterfeit products

Illicit drugs

harmful content

disinformation

2020 *This Is What Happens In An Internet Minute*



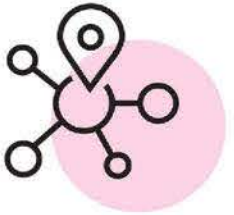
In a nutshell the objectives are:



Modernise the rules to **more effectively** address **illegal content** and **systemic risks** in the online space



Clarify the **rules on liability**, giving companies legal certainty to take voluntary action **in a diligent and proportionate manner under clear terms of service**



Increase transparency, accountability and facilitate **better oversight**

Intermediary services

- **offering network infrastructure: Internet access providers, domain name registries, wifi hotspots...**

Hosting services

- **such as cloud infrastructure and webhosting services**

Online platforms

- **E.g. online marketplaces, app stores, or collaborative economy platforms or social media platforms**

Very large online platforms

- **Specific rules for platforms reaching 45 million users (10% of EU population)**

Liability provisions

DSA harmonises
liability
exemptions, not
liability

Maintains the principles of the **eCommerce directive**

Does not attribute liability. The contrary would mean to harmonise tort or criminal law

Maintains the prohibition on general monitoring.

DSA is **neutral**

Does not define what is illegal

Illegality is defined by national or EU rules

DSA is **horizontal**

All types of **illegal content**, civil and criminal liability

Some provisions can also cover **harmful content** (transparency & safeguards for decisions under terms of service, systemic risks on very large platforms)

DSA **applies to services**, not to actors

Not *"Platform 'X' is an intermediary"*, but *"'X' provides intermediary services"*

A provider can provide many different services (mere conduit + caching + hosting services which are not intermediary services or not even ISS)

All obligations

OBLIGATIONS	VERY LARGE PLATFORMS	ONLINE PLATFORMS	HOSTING SERVICES	ALL INTERMEDIARIES
Points of contact	•	•	•	•
Legal representatives	•	•	•	•
Terms and conditions	•	•	•	•
Transparency reporting	•	•	•	•
Notice & Action	•	•	•	
Statement of reasons	•	•	•	
Complaint handling	•	•		
Out of Court Dispute Settlement	•	•		
Trusted flaggers	•	•		
Abusive behaviour	•	•		
Know Your Business Customer (KYBC)	•	•		
Reporting criminal offences	•	•		
Advertising transparency	•	•		
Additional transparency reporting	•	•		
Risk assessment and mitigation	•			
Independent audits	•			
Recommender systems	•			
Enhanced advertising transparency	•			
Data access and scrutiny	•			
Compliance officer	•			
Enhanced Transparency reporting	•			

Cumulative obligations

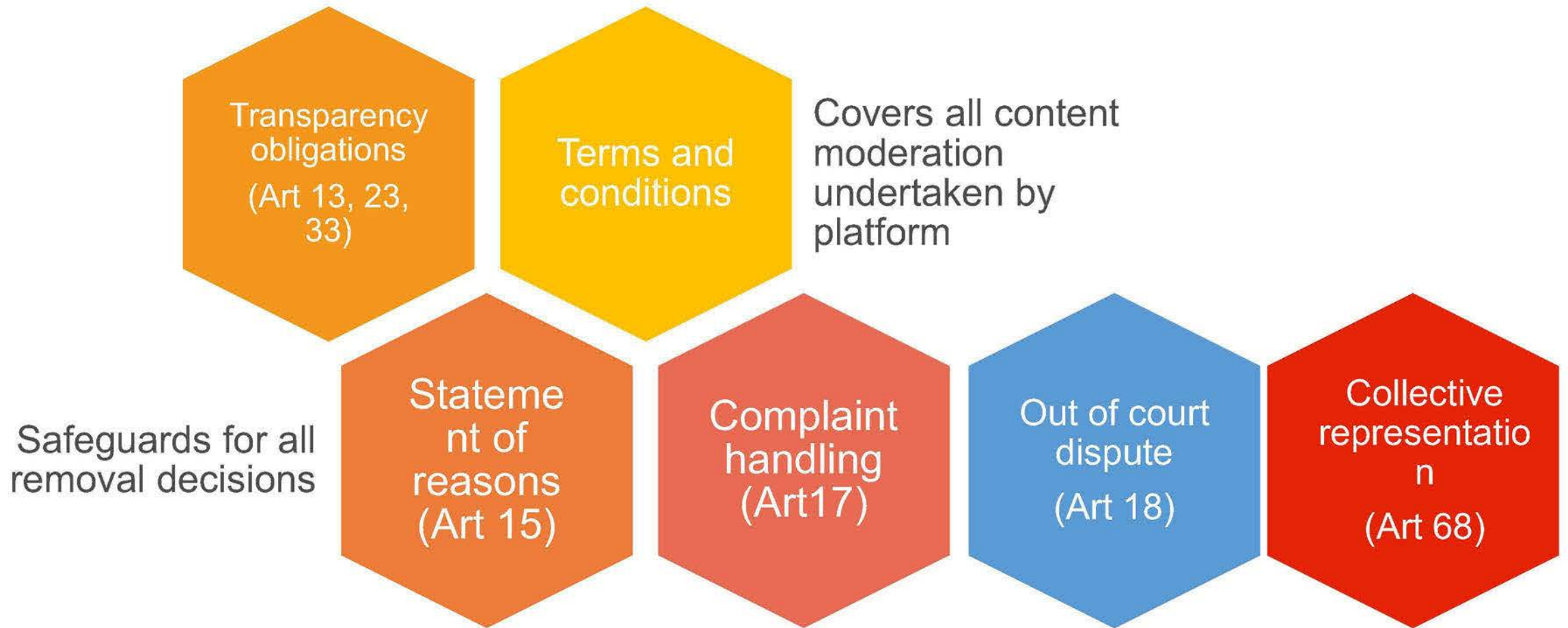
Effective measures for addressing illegal content

Orders
(Art 8)

Notice
& Action
(Art 14)

Trusted
flaggers
(Art 19)

Due process & transparency

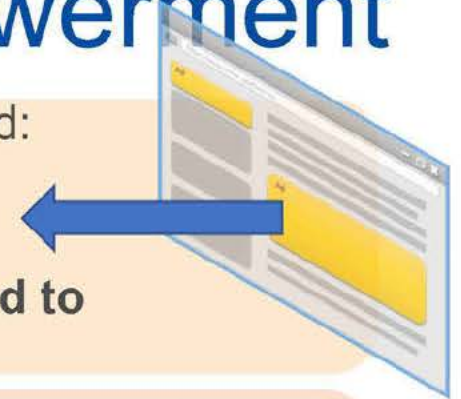


Addressing amplification & user empowerment

Advertising transparency

Art 24

- Ensure that users can identify **in real time** when seeing an ad:
 - **That the information is an advertisement**
 - **The advertiser**
 - **Why the ad is displayed for them: main parameters used to determine this.**



Ad repositories

Art 30

- Publicly available **repository** including for all ads displayed on the platform
 - **The advertisement & advertiser**
 - Period during which the ad was displayed
 - **Whether it was a targeted ad & targeting criteria**
 - Total number of users reached & aggregate data on delivery



Recommender systems

Art 29

- Main parameters in ToS
- Include an option for users which is not based on profiling

Binding risk management obligations on very large online platforms



Co-regulatory framework

Codes of Conduct Art 35

- Codes of conduct to be supervised by the Commission and the Board
- Can serve as compliance measures to fulfill the obligations under Article 27.

Crisis Protocols Art 37

- Addressing crisis situations strictly limited to extraordinary circumstances affecting public security or public health
- Limited in time, clear procedures and roles, safeguards and public reporting.

Governance structure

Digital Services Coordinators

- Independent authorities
- Direct supervision and enforcement (by default)
- Coordination with other national competent authorities

European Board for Digital Services

- Ad-hoc independent advisory group
- Composed by national Digital Services Coordinators
- Chaired by COM
- Advising DSCs and COM, recommending actions

European Commission

- Direct enforcement powers vis à vis very large online platforms for
- Advises on cross-border disputes
- Intervenes upon DSC request

Breaches of DSA obligations => **Digital Services Coordinators** and **COM** can impose various sanctions and fines up to **6%** of the provider's turnover

Subgroup on combating illegal hate speech online

Referral Action Day against Right-Wing Terrorist Online Propaganda

Redacted

EUROPOL

A photograph of the Europol building in The Hague, a modern multi-story structure with a grey facade and large windows. The building is partially obscured by a blue overlay on the right side of the image. In the top left corner, there is a white circular graphic with the Europol logo inside. The text 'EUROPOL' is visible on the building's facade and in the bottom left corner.

EUROPOL

First RWTCO Referral Action Days 2021

07/07/2021
EU INTERNET REFERRAL UNIT
European Counter Terrorism Centre

EUROPOL

Europol Unclassified – Basic Protection Level

Objective

Overview - Understanding the problem and positions of MSs and OSPs

Participants

18 EU-MS & 10 Third parties

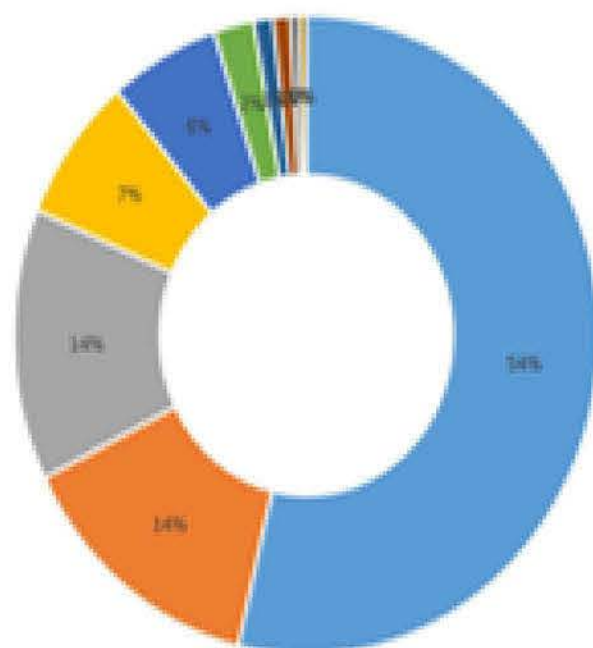
Concept

Collection - Deconfliction - Referral

Outcome

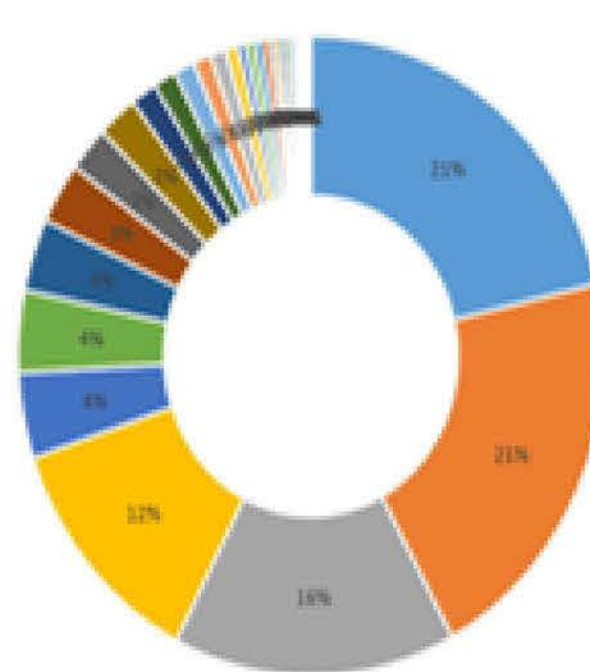
1038 URLs, spread over well-known & 'new' OSP's (forums, websites, blogs)

URLs referred by contributor



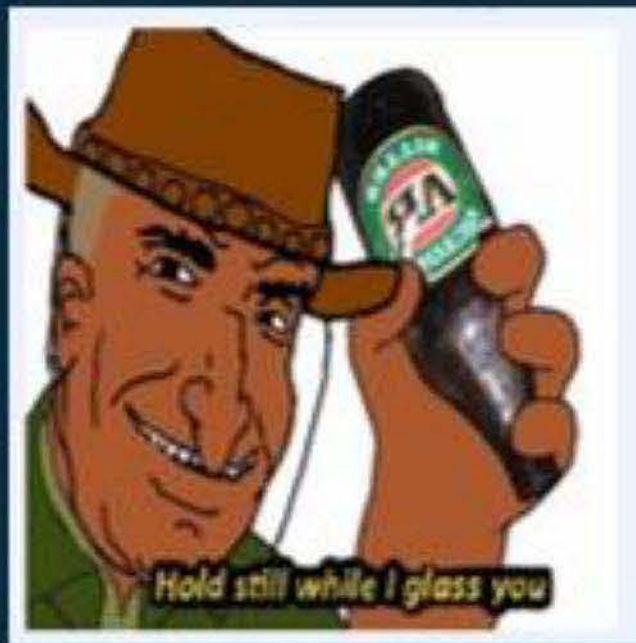
- Germany
- Belgium
- Europe
- Italy
- Luxembourg
- Slovakia
- United Kingdom
- Czech Rep
- Australia
- Spain

URLs referred by platform



- VK
- Unregistered Hosting Party
- Telegram
- Archiv.org
- Drop and .com
- Bitchute.com
- Blogspot
- Gali.com
- Meerkat.com
- Twitter
- Reddit
- Facebook
- YouTube
- Instagram
- Wordpress
- Freegoff.info
- Gambia

Outcome





Thank you

07/07/2021
EU INTERNET REFERRAL UNIT
European Counter Terrorism Centre



FRA-led working group on hate crime recording, data collection, and encouraging reporting

Encouraging hate crime reporting: the role of law enforcement and other authorities

Redacted

Rights

European Union Agency for Fundamental

ENCOURAGING HATE CRIME REPORTING

THE ROLE OF LAW ENFORCEMENT AND OTHER AUTHORITIES

REPORT



1	ENCOURAGING HATE CRIME REPORTING: DELIVERING ON THE LEGAL DUTY TO FULLY RESPECT, PROTECT AND PROMOTE FUNDAMENTAL RIGHTS	
1.1	DUTY TO ENSURE ACCESS TO JUSTICE FOR ALL	
1.2	DUTY TO PROTECT AND SUPPORT VICTIMS OF HATE CRIME	
1.3	DUTY TO INVESTIGATE AND PUNISH HATE CRIME	
1.4	INTERNATIONAL AND EUROPEAN POLICY GUIDANCE ON ENCOURAGING REPORTING	
	ENDNOTES	
2	UNCOVERING (UNDER)REPORTING	
2.1	BIAS-MOTIVATED VIOLENCE AND HARASSMENT: AN ALL TOO COMMON OCCURRENCE IN THE EU	
2.2	UNDERREPORTING HINDERS VISIBILITY OF VICTIMISATION	
2.3	TO WHOM DO VICTIMS REPORT?	
2.4	REASONS FOR NOT REPORTING	
2.5	PARTICULAR CHALLENGES IN REPORTING FACED BY CERTAIN GROUPS AT RISK OF HATE VICTIMISATION	
	ENDNOTES	
3	ENCOURAGING HATE CRIME REPORTING: ENABLING ENVIRONMENT, STRUCTURES AND PROCESSES	
3.1	CULTIVATING AN ENABLING SOCIAL ENVIRONMENT	
3.2	ENABLING STRUCTURAL FACTORS: DIVERSE REPORTING OPTIONS AND STANDARDISED REFERRALS	
3.3	ENABLING PROCESS FACTORS: BUILDING INSTITUTIONAL CAPACITY AND STRENGTHENING PARTNERSHIPS	
	ENDNOTES	
4	WAYS FORWARD: TOWARDS A VICTIM-CENTRED APPROACH TO HATE CRIME REPORTING	

Reporting - a pre-condition for meeting legal commitments

When hate crimes are not reported:

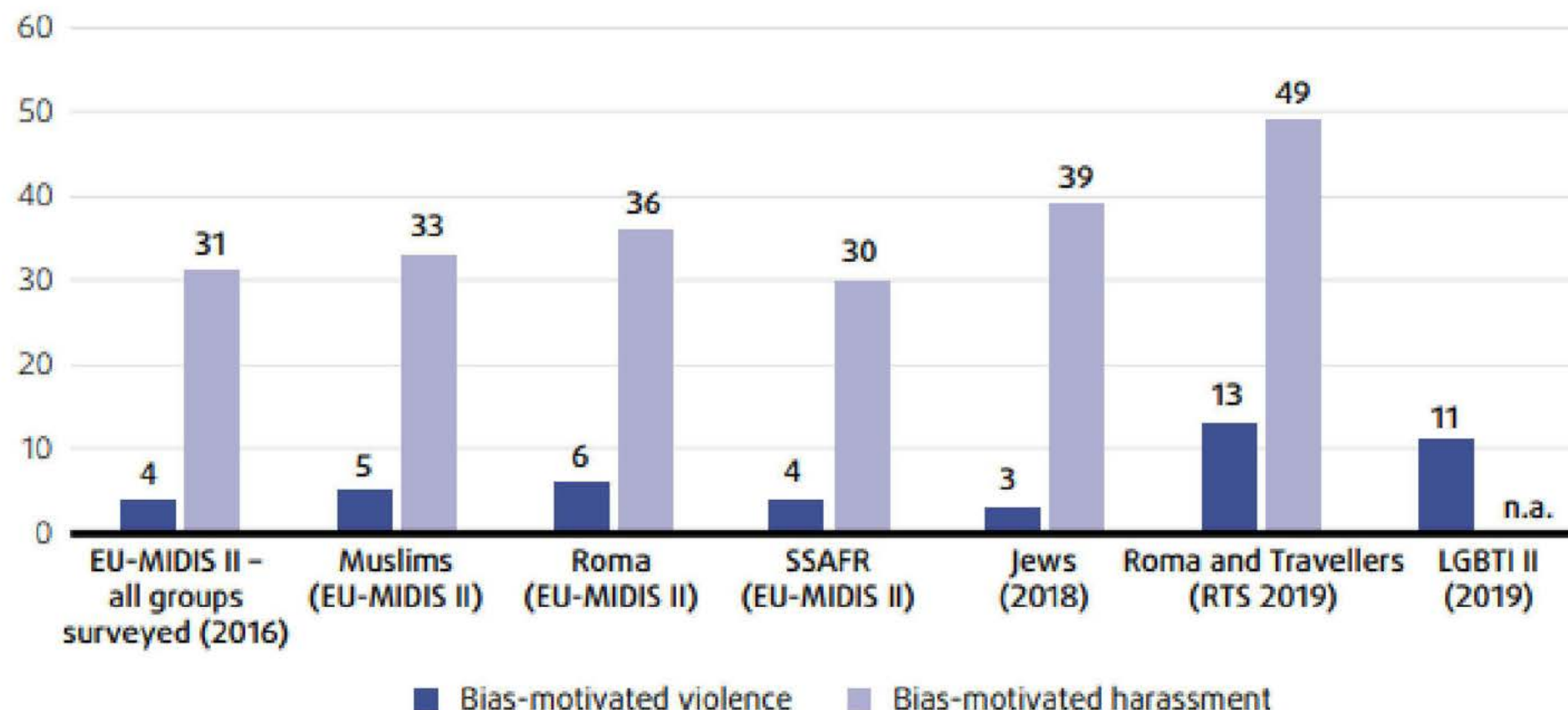
- victims do not get the support they need;
- their access to justice is limited;
- perpetrators remain unpunished; and
- law enforcement and criminal justice authorities are not able to respond appropriately to hate crime.

Member States can not deliver on their duties to:

- ensure access to justice for all,
- protect and support victims, and
- combat hate crime.

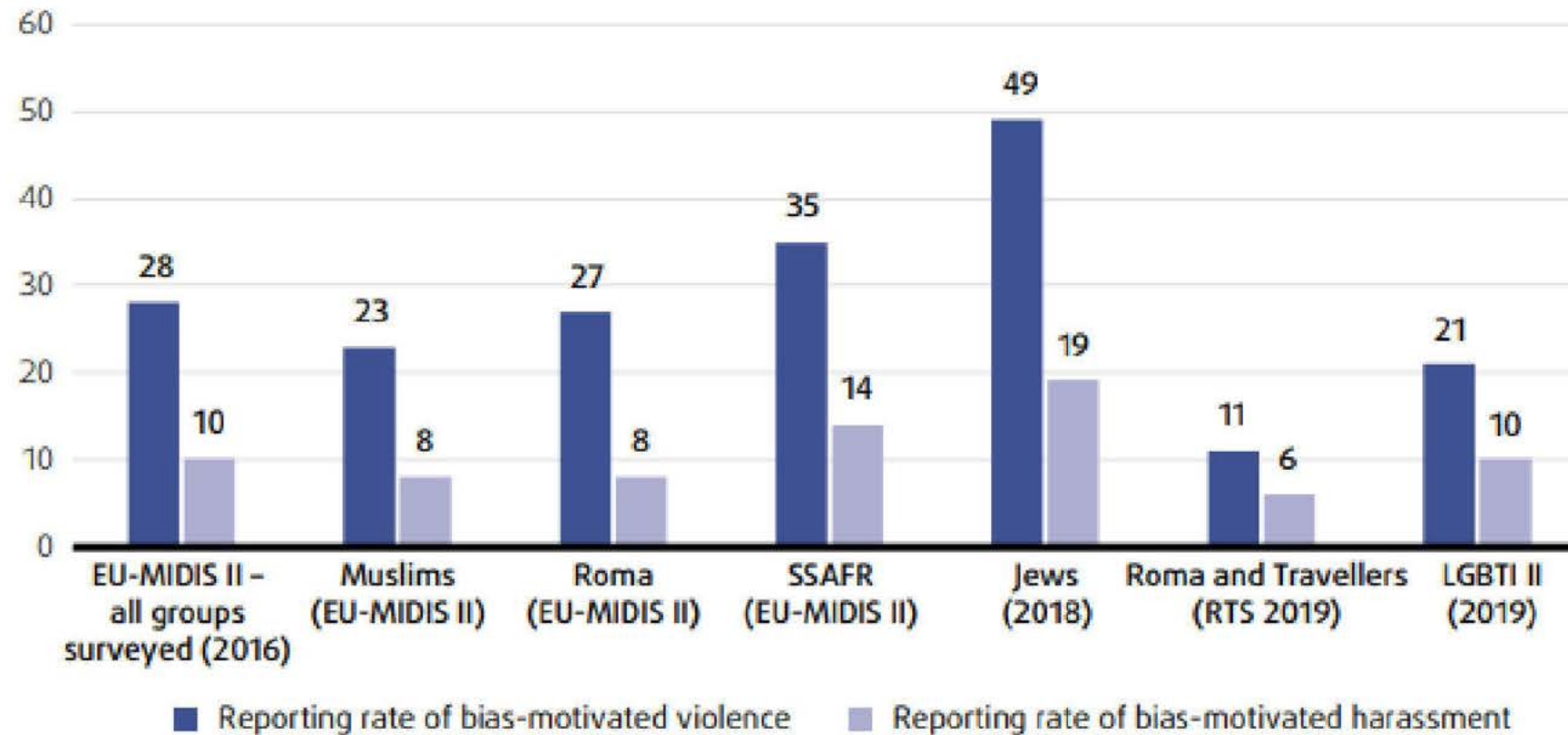
Uncovering (under)reporting

FIGURE 1: PREVALENCE OF BIAS-MOTIVATED VIOLENCE AND HARASSMENT IN THE FIVE YEARS BEFORE THE FRA SURVEY IN QUESTION, BY SURVEY GROUP (%)^{a,b,c,d,e,f}



Sources: FRA, EU-MIDIS II, 2016; second Survey on Discrimination and Hate Crime against Jews in the EU, 2018; Roma and Travellers Survey, 2019; and EU LGBTI Survey II, 2019

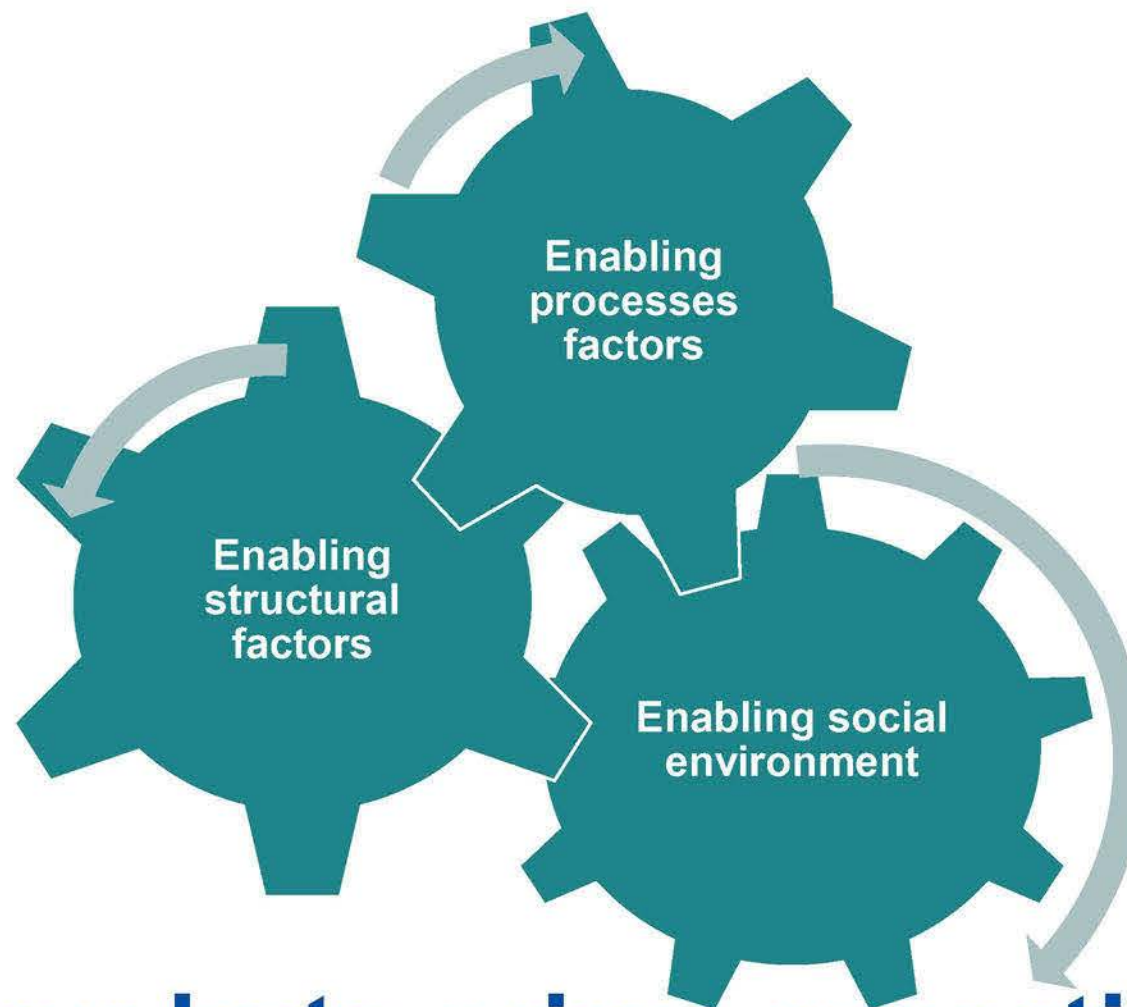
FIGURE 3: REPORTING OF THE MOST RECENT INCIDENT OF BIAS-MOTIVATED VIOLENCE AND HARASSMENT TO THE POLICE OR ANY OTHER ORGANISATION IN THE FIVE YEARS BEFORE THE FRA SURVEY IN QUESTION, BY SURVEY GROUP (%)^a



Sources: FRA, EU-MIDIS II, 2016; second Survey on Discrimination and Hate Crime against Jews in the EU, 2018; Roma and Travellers Survey, 2019; and EU LGBTI Survey II, 2019

Reasons for not reporting

- Incidents commonplace, little hope for change
'Nothing will change' & 'It happens all the time'
- Cumbersome reporting procedures
'Too bureaucratic and time consuming'
- Experiences erode trust in police
'I don't trust the police'



Encouraging hate crime reporting: drivers and barriers

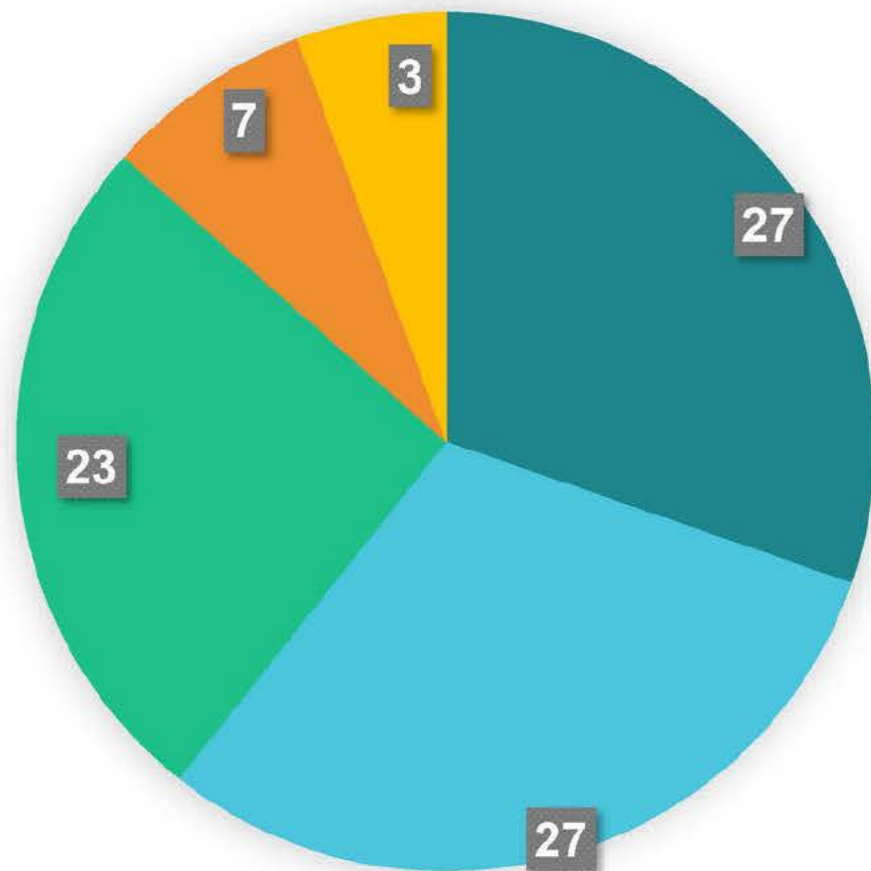
Enabling social environment

- Countering discrimination by law enforcement authorities
- Police leadership condemns hate crimes and communicates publically about hate crime cases
- Create partnerships and regularly engage with communities and CSOs
- Provide accessible and inclusive information and support services
- Raise awareness about rights, victim support services and existing complaint mechanisms

Enabling structural factors

- Alternative reporting options
 - Third party reporting and anonymous reporting
- Structured referrals between police, equality bodies, third party reporting centers, CSOs and victims support services
- Reported incidents are correctly recorded and data collected

Reporting options

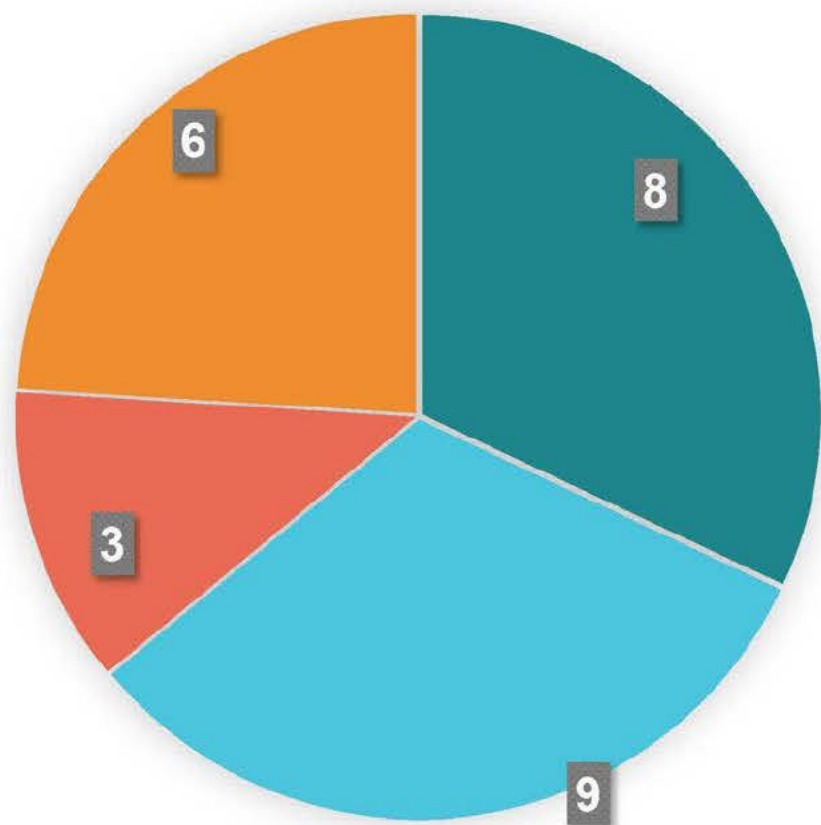


Total reporting options in (EU 26 and UK)

- In person
- Phone line / hotline (24h)
- Online form
- Social media networks
- Live web chat

*Sources: Data provided by members of FRA's WG in 2020 - 26 EUMSs and UK

Structured mechanism for referrals of hate crime victims to victims support services



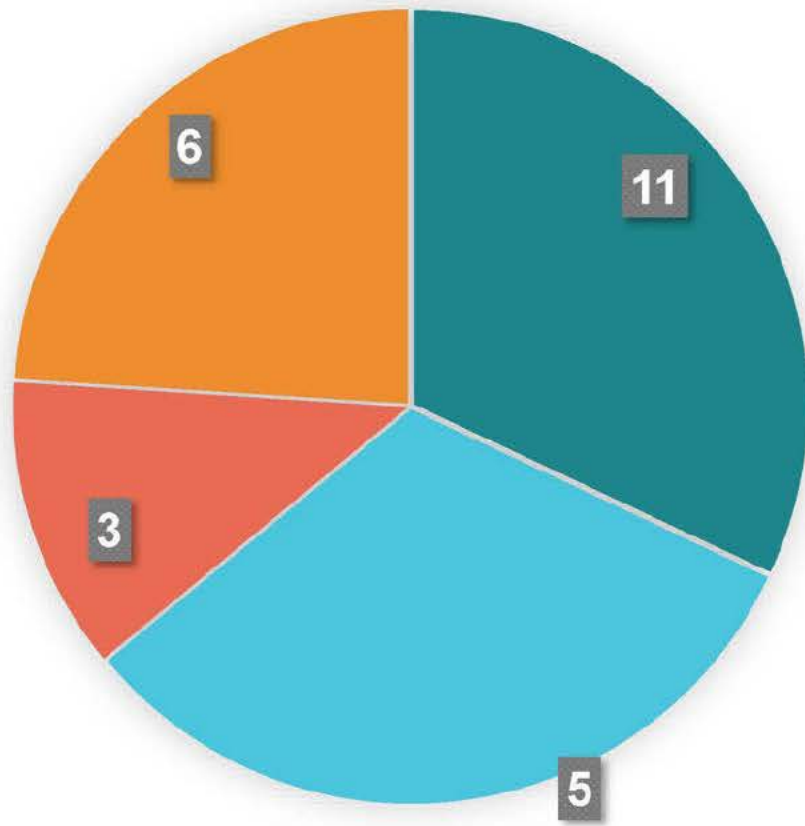
Referral mechanisms (EU 26 and UK)

- No referral mechanism
- Police to/from VSS
- Police to VSS
- VSS to Police

*Sources: Data provided by members of FRA's WG in 2020 - 26 EUMSs and UK

Structured mechanism for referrals of hate crime victims to and from Equality bodies

Referral mechanisms (EU 26 and UK)



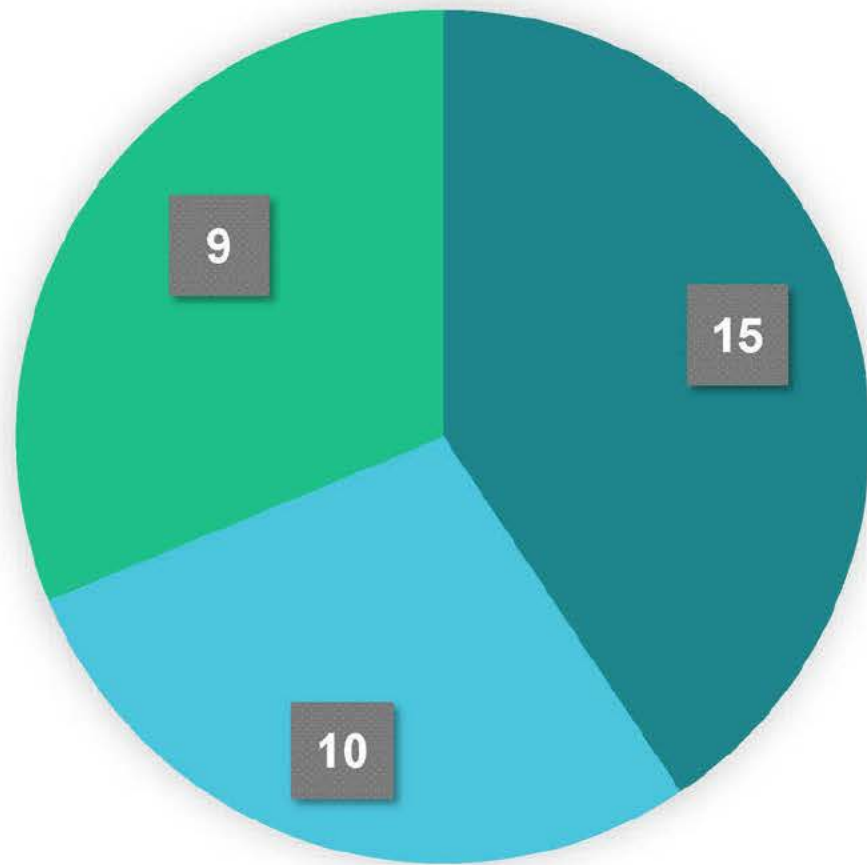
- No referral mechanism
- Police to/from EB
- Police to EB
- EB to Police

*Sources: Data provided by members of FRA's WG in 2020 - 26 EUMSs and UK

Enabling process factors

- Practical police guidance
- Specialisation - specific hate crime officers and dedicated hate crime units
- Police training and capacity building
- Cooperation and coordination
 - coordination between public services
 - coordination with other organisations, beyond public bodies - VSSs and CSOs

Hate crime specialisation



Specialists or units for hate crime in
26 Member States and UK)

- Specialist hate crime officer
- Hate crime unit
- None

Thank you

hatecrime@fra.europa.eu

OSCE/ODIHR Expert Network for Hate Crime Victim Support (EStAR project)

Project presentation

Redacted

OSCE/ODIHR

Redacted

Association of Counselling Centers for Victims of
Right-wing, Racist and Anti-Semitic Violence in Germany (VBRG)



Enhancing Stakeholder Awareness and Resources for Hate Crime Victim Support - EStAR -

The 9th Meeting of the EU High-Level Group on Combating Racism, Xenophobia and Other Forms of Intolerance

7 July 2021

Presented by: **Redacted** OSCE/ODIHR
Redacted, VBRG

Milestones reached

- **Network of Experts:** Enhanced engagement
- **Diagnostic Tool Testing Phase Began:** Pilot diagnostic workshops to assess national structures for hate crime victim support
- **Publication:** Model Quality Standards for Specialist Support
- **Publication:** Model Guidance on Individual Needs Assessments

EStAR Tools

a comprehensive approach to supporting hate crime victims

- Understanding and responding to victim's needs
- Diagnostic of national hate crime victim support structures and addressing the gaps
- Sensitive and respectful treatment
- Providing specialist support services
- Ensuring CSO-Government collaboration in providing protection and support.



1113



European
Commission

Network of Experts: Enhanced engagement

"Under the Spotlight" Series

EStAR discussion series showcasing current and emerging topics related to hate crime and victim support to provide interactive discussion and networking space with expanded stakeholders beyond the working group.

First talk: Understanding the needs of hate crime victims with disabilities.

Hate Crime Training for Polish Civil Society

Presentation of EStAR's Understanding the Needs of Hate Crime Victims

Discussed strategies for combating and monitoring hate crimes and how to enhance hate crime victim support structures.

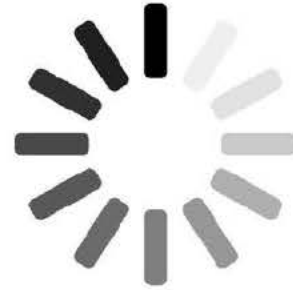


Eastern Partnership Peer Meeting on Victim Support Services

EStAR project presented its work during an online discussion on hosted by the Council of Europe on "Victim support services for victims of hate speech / hate crimes and making hate crime investigation more victim-centred"

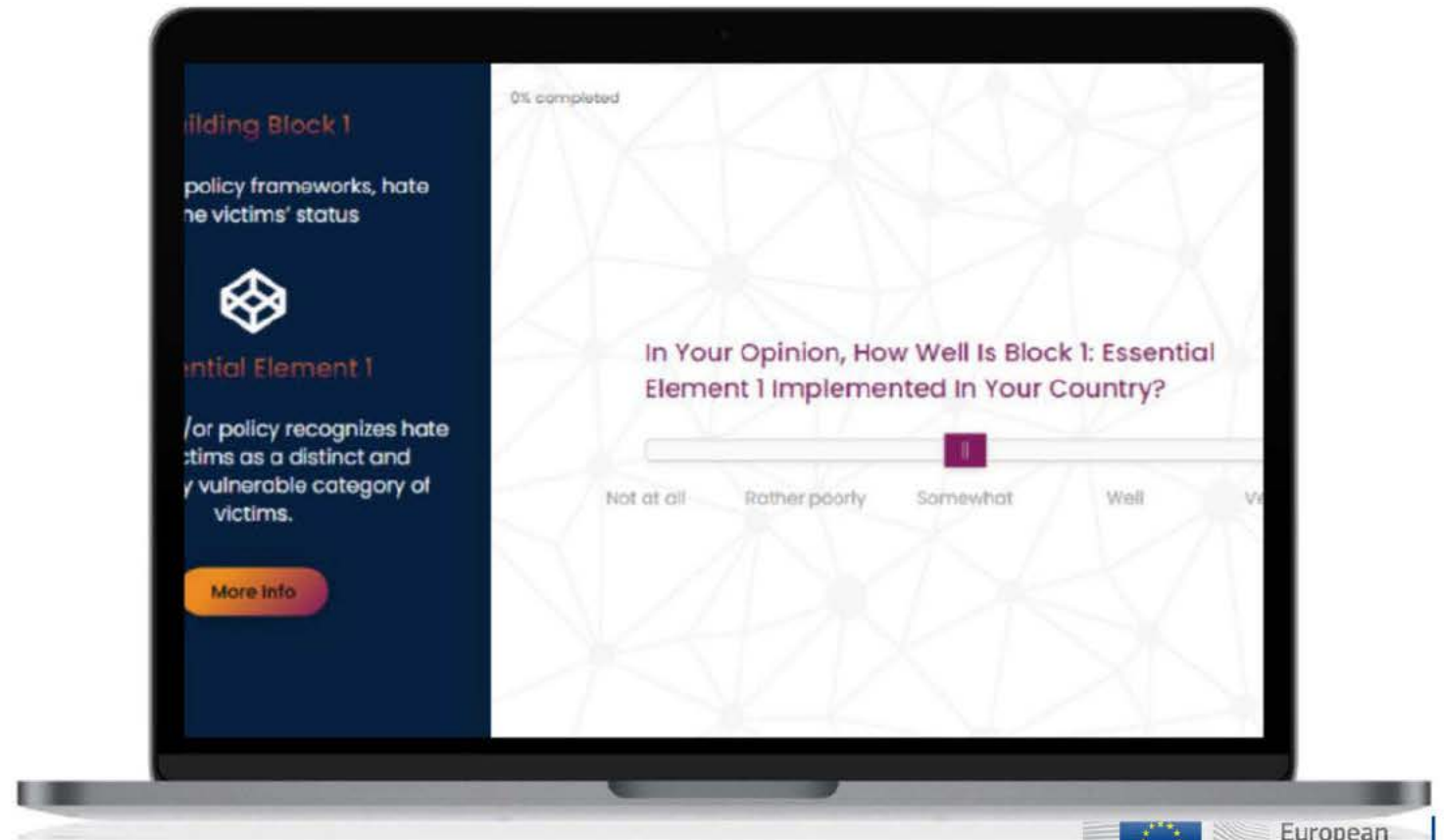


Diagnostic Publication

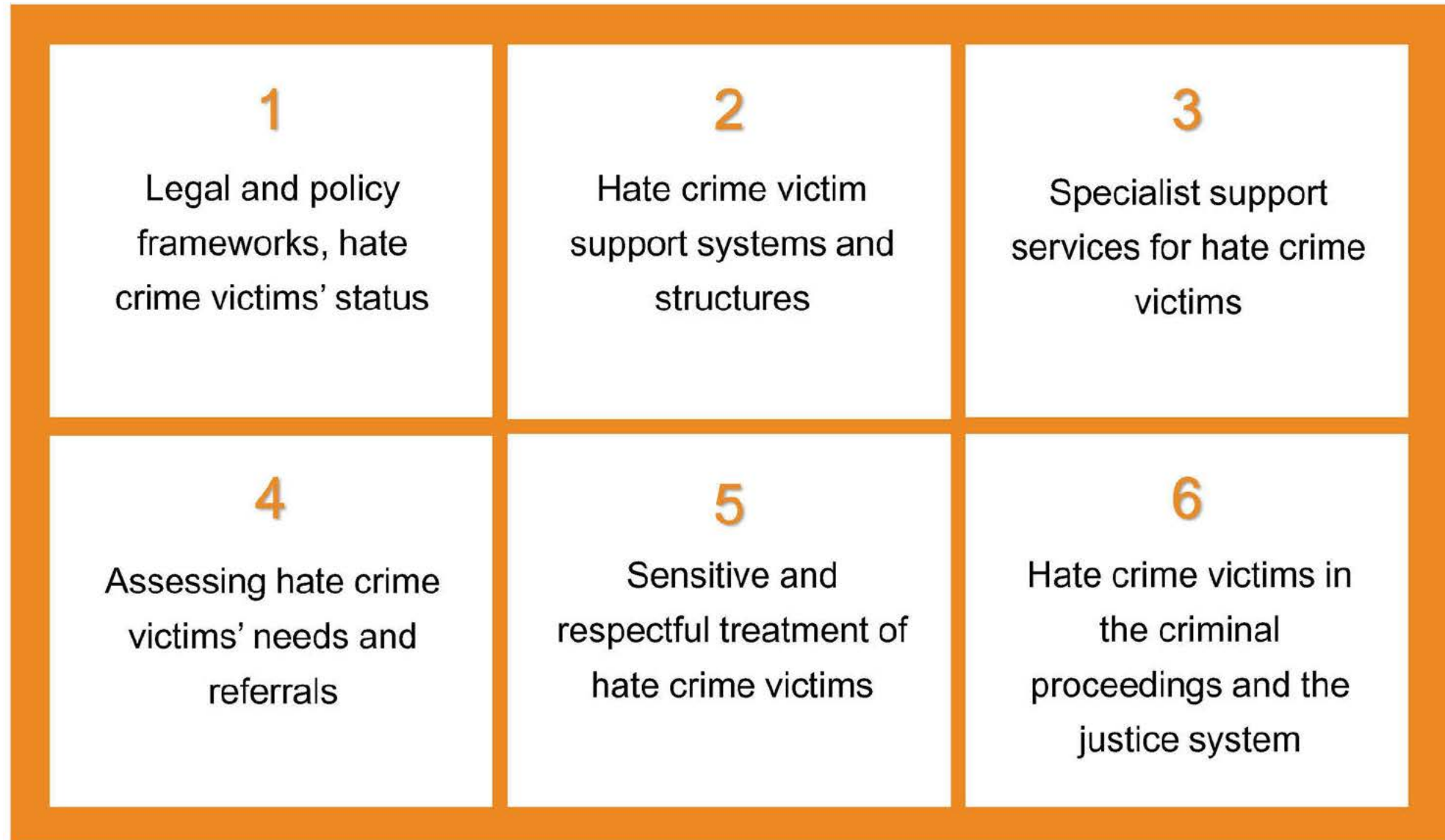


Digital Diagnostic Tool

Resources to assist hate crime victim support practitioners, officials, activists and policy makers to identify strengths and gaps in their national systems, in order to understand how best to address shortcomings and build on opportunities to enhance protection and support for victims of hate crime.

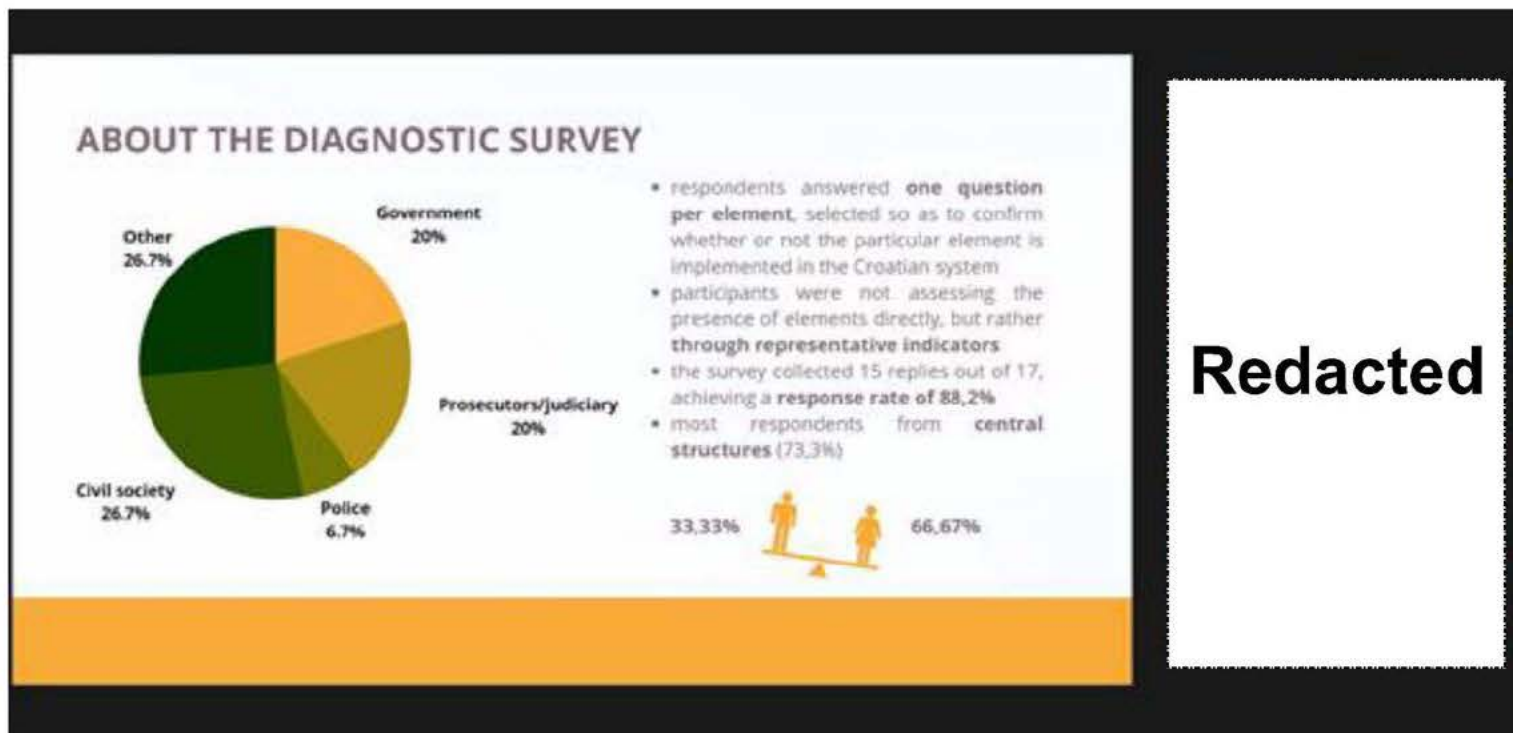


Building Blocks



Piloting the Diagnostic Tool

Participants found the workshop interactive and relevant to their work, appreciated clear actionable results it produced and the fact that it brought together various stakeholders (particularly government and CSOs).



1st Workshop in Spain

30 March 2021

2nd Workshop in Croatia

16 June 2021

3rd Workshop

In discussion - to be confirmed shortly

4th Workshop



Forthcoming resources and tools:

Sensitive & Respectful Treatment



Model Guidance on Sensitive and Respectful Treatment of Hate Crime Victims (publication)



Training Course on Sensitive and Respectful Treatment of Hate Crime Victims for law enforcement and criminal justice professionals (prosecutors and the Judiciary)



National pilot training in Portugal

In collaboration with the Centre for Judicial Studies (Centro de Estudos Judiciários)



“

Model Quality Standards for Hate Crime Victim Support

Quality standards play a crucial role in developing and strengthening specialized hate crime victim support, as high professional standards for the conduct, expertise and organization of services are key to meeting victims' needs.



Model Quality
Standards
for Hate Crime
Victim Support



Publication Highlights



- The guide promotes a victim-centered approach to the development and delivery of support services to hate crime victims
- The guide recognizes the direct and indirect impact of hate crimes on individuals, communities and society
- The guide stresses the need for adequate provision of resources to service providers
- The guide emphasizes cooperation and effective referrals.

Next steps

Training Course for Specialist Support Providers

Pilot regional training for specialist support providers

“

Model Guidance on Individual Needs Assessments (INA)

The INA is a key instrument to ensure a victim-centred approach. It is the first step to protecting victims' rights and ensuring that all necessary services are provided - as required by the Victims' Rights Directive.



Model Guidance
on Individual Needs
Assessments of
Hate Crime Victims

Publication Highlights & INA Essentials

- Needs of hate crime victims require additional considerations – the guide promotes a two-step approach that requires to carry out an in-depth INA
- Effective implementation of INA should be accompanied with additional resources, training and operational guidelines
- Cooperation between all relevant stakeholders is key to a sensitive and effective INA pr
- Proactive approach is required to community needs.

INA Good Practice

Virtual study visit to France
19-20 May 2021

Offered insights into the system France has established to assess and address the needs of hate crime victims. The study visit illustrated how criminal justice agencies and victim support organizations cooperate in assessing needs, providing support services, protection and access to justice.



Next steps

Compendium on frameworks for INA

Provides information on how INAs are conducted in various project countries and takes stock of the existing encouraging practices and initiatives.

Contact information

Redacted OSCE/ODIHR: **Redacted** @odihr.pl

Redacted VBRG: **Redacted** @verband-brg.de

Project website:

<https://www.osce.org/odihr/hate-crime-victim-support>



ASSOCIATION OF COUNSELING CENTERS
FOR VICTIMS OF RIGHT-WING, RACIST AND
ANTI-SEMITIC VIOLENCE IN GERMANY



European
Commission

CEPOL-led working group on hate crime training for law enforcement

Redacted

CEPOL

Working Group on Hate Crime Training

**HLG on Combating Racism,
Xenophobia and other Forms of
Intolerance**

07 July 2021

**EDUCATE
INNOVATE
MOTIVATE**

EUROPEAN UNION AGENCY FOR LAW ENFORCEMENT TRAINING

CEPOL - HATE CRIME TRAINING 2021

Onsite Training Activity on Hate Crime
Sapienza University
Rome
16-19 November 2021

Aim: To enhance the police response to hate crime in general and online hate speech in particular as a specific form of delinquency with special motives, as well as identifiable social and psychological basis and to encourage the use of specific investigation methods and law enforcement cooperation in the field of hate crime.

Target Group

Law enforcement officials and specialists responsible of tackling hate crime and online hate speech.



CEPOL - HATE CRIME TRAINING 2021

❖ 5 Webinars on Hate Crime issues

LGBTQ Community	ILGA- Europe + Háttér Society
Anti-Muslim Hatred	UK Expert Redacted
Anti-Semitism	European Jewish Congress
Roma and Travellers	Fundamental Rights Agency
Third Party Reporting	Fundamental Rights Agency, RIAS Berlin, COMPAS Charity

CEPOL - HATE CRIME TRAINING 2021

❖ E-Lesson on Anti-Muslim Hatred

- Islam and Muslims in Context
- Definitions of Islamophobia and Anti-Muslim Hatred
- Recognising Anti-Muslim Hatred-driven Hate Crime
- Barriers to reporting
- Supporting Victims
- Case Studies and Best Practice



Expert Group on Fundamental Rights

- Consists of 9 MS (AT, ES, EL, IE, IT, LV, PL, PT, SI)
- COM, EIGE, EUROPOL and FRA
- Co-chaired by CEPOL Executive Director and expert from AT
- 2 meetings: 26 March and 29 June 2021



Expert Group on Fundamental Rights

3 Sub-groups:

- 1) Overview existing and new relevant international and EU strategic documents and instruments on FR and FR training Co-chaired by CEPOL;
- 2) Analysis of the status of FR training in CEPOL and identify gaps and shortcomings;
- 3) Overview of current partnerships, including such outside of law enforcement, and identify enhancement opportunities for cooperation.

=> Strategy Document on FR training for LE



WORKING GROUP ON HATE CRIME TRAINING

- ✓ 1st meeting on 24-25 March 2021
- ✓ 80 Attendees from 23 MS and 14 CSOs and the FRA, Council of Europe and OSCE/ODIHR
- ✓ Mandate of the Working Group
- ✓ Overview of hate crime training and support for LE by CEPOL, FRA, ODIHR and the Council of Europe
- ✓ Mapping Study of Hate Crime Training for LE and judicial professionals in Europe – CEJI
- ✓ Way forward



Mandate of the WG

- EU Anti-Racism Action Plan and the Anti-Racism Summit
- Support the implementation of the 10 key guiding principles
- Support the MS in their duty to ensure full transposition of EU Law
- Advancement of adequate training for LE and judicial staff in the EU MS
=> Improve effective investigation of HC, promote working in partnerships and apply the law in an effective manner

Overview COM, CEPOL, FRA, Council of Europe and ODIHR

- EU Anti-Racism Action Plan
- CEPOL Training and Expert Group on Fundamental Rights
- CoE Human Rights Education for Legal Professionals
- ODIHR Training Programmes for LE and Prosecutors (TAHCLE and PAHCT)



Mapping Outcomes

- Comprehensive Mapping Exercise done by CEJI as an **essential tool** for the WG's work;
- Provided a general **state of play** of the training activities offered at national, regional and local levels;
- Conducted an **assessment of needs and gaps** based on the 10 key guiding principles

Methodology

- 1) Desk Research
- 2) Questionnaire
- 3) Analysis of the information and follow up
- 4) Draft Report

Results

1. 25 questionnaires received
2. 16 questionnaires provided information on both law enforcement and criminal justice agencies
3. Example of existing practices
4. Qualitative approach based on the information submitted by the national authorities;

Influencing factors

1. Hate Crime legislation available
2. Hate Crime training is integrated in strategies/policies
3. Level of reporting and recording
4. Use of resources/cooperation at European Level (OSCE/ODIHR, CEPOL, CEJI's Facing Facts, EJTN, EU projects)

Central components of hate crime training

- **Ensuring impact and sustainability**
 - Principle One: Ensuring impact and sustainability – embed training within a broader approach to tackling hate crimes
- **Assessing training needs and identifying targets**
 - Principle Two: Develop a methodology to assess training needs
 - Principle Three: Customise programmes on the basis of the identification of target personnel
- **Building synergies with civil society organisations**
 - Principle Four: Develop a model of structured cooperation with civil society
- **Choosing the right methodology**
 - Principle Five: Combine different training methodologies, including practice-based sessions
 - Principle Six: Consider the development of train-the-trainer programmes
- **Conveying quality content**
 - Principle Seven: Focus on achieving targeted overall objectives through quality content
 - Principle Eight: Develop targeted training modules to address specific forms of intolerance
- **Monitoring and evaluating outcomes**
 - Principle Nine: Link hate crime training to mainstream performance review processes
 - Principle Ten: Ensure regular monitoring and evaluation of training outcomes

Conclusions

- P1: Support required for a national/regional strategic approach to planning sustainable and regular HC training
- P2+P3: Promote a more general and centralised TNA and support the development of systematic TNA methodology
- P4: a more structured engagement of CSOs
- P5: blended learning is generally given
- P6: Enhance the use of the Train-the-Trainers method for cascading

Conclusions

- P7: Raise the level of training beyond awareness raising and knowledge increase;
- P8: Support targeted training modules on specific forms of intolerance;
- P9: Enhance the consideration of HC training in individual performance review processes;
- P10: Reinforce the evaluation of mid- and long-term training outcomes

WORKING GROUP ON HATE CRIME TRAINING

- ✓ MS to nominate experts;
- ✓ Next meeting: to focus on the main gaps encountered in the mapping, including methodology for assess training needs and identify targets

Thank you!

European Union Agency for Law Enforcement Training

Offices: H-1066 Budapest, Ó utca 27., Hungary • Correspondence: H-1903 Budapest, Pf. 314, Hungary

Telephone: +36 1 803 8030 • Fax: +36 1 803 8032 • E-mail: info@cepol.europa.eu • www.cepol.europa.eu



The Citizens, Equality, Rights and Values programme and Commission's funding opportunities

Redacted

European Commission

Overview

1. Citizens, Equality, Rights and Values programme
2. Work Programme 2021-2022
3. How to apply?

The multi-annual financial framework (MFF) 2021-2027



European Parliament and Council agreement of 10 November 2020

Parliament and EU leaders reached a **historic agreement on the next-long term budget and the recovery instrument NextGenerationEU:**

- Largest package ever financed through the EU budget, of **€ 1.8 trillion**.
- For the first time, the EU will have a **specific mechanism** to protect its budget against breaches to the **rule of law**.
- **Ten programmes received top-ups** compared to the European Council agreement from 21 July 2020.

Long-term budget 2021-2027

Allocations per heading

	MFF	NEXT GEN. EU	TOTAL
1. Single Market, Innovation and Digital	132.8	10.6	143.4
2. Cohesion, Resilience and Values	377.8	721.9	1 099.7
3. Natural Resources and Environment	356.4	17.5	373.9
4. Migration and Border Management	22.7	-	22.7
5. Security and Defence	13.2	-	13.2
6. Neighbourhood and the World	98.4	-	98.4
7. European Public Administration	73.1	-	73.1
TOTAL MFF	1 074.3	750.0	1 824.3

All amounts in EUR billion, in 2018 prices

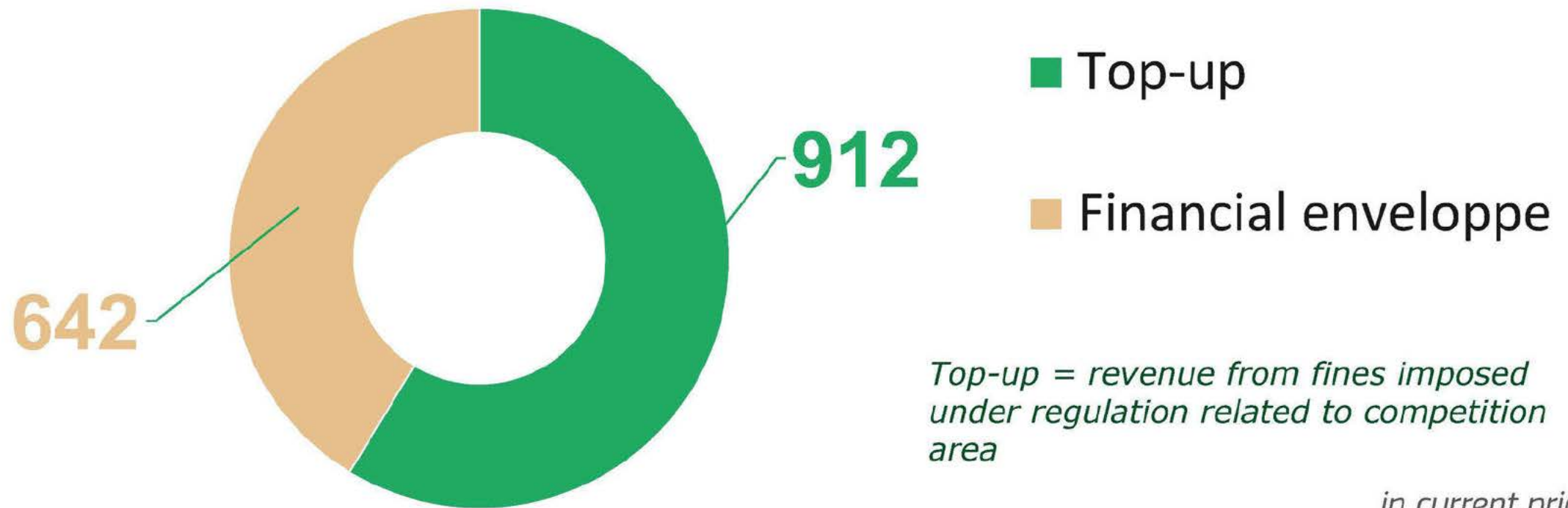
€ 15 billion top-up explained



All amounts in EUR billion, 2018 prices.
Source: European Commission.

Citizens, Equality, Rights and Values

Budget: € 1.5 billion



Top-up = revenue from fines imposed under regulation related to competition area

in current prices



EU funding opportunities

Citizens, Equality Rights and Values programme (CERV)

Merge of two existing programmes :

- the Rights, Equality and Citizenship programme
- the Europe for Citizens programme

Justice programme

Financial programmes

key instruments for the implementation of EU policies



- Treaties (and Charter)
- Legal instruments
- Financial Programmes

Citizens, Equality, Rights and Values programme



Objective: to protect and promote rights and values as enshrined in the Treaties, to sustain and further develop rights-based, democratic, equal and inclusive societies based on the rule of law.

[Regulation \(EU\) 2021/692](#): **EUR 1.5 billion** for the 2021-2027

What are « EU values » ?



Article 2 of the Treaty on the European Union

The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.

Legal basis

Treaty on the Functioning of the European Union

Article 16(2)	Data protection
Article 19(2)	Non-discrimination
Article 21(2)	Citizens' rights to move and reside freely
Article 24	European citizens' initiative
Article 167	Culture
Article 168	Health

Areas of intervention

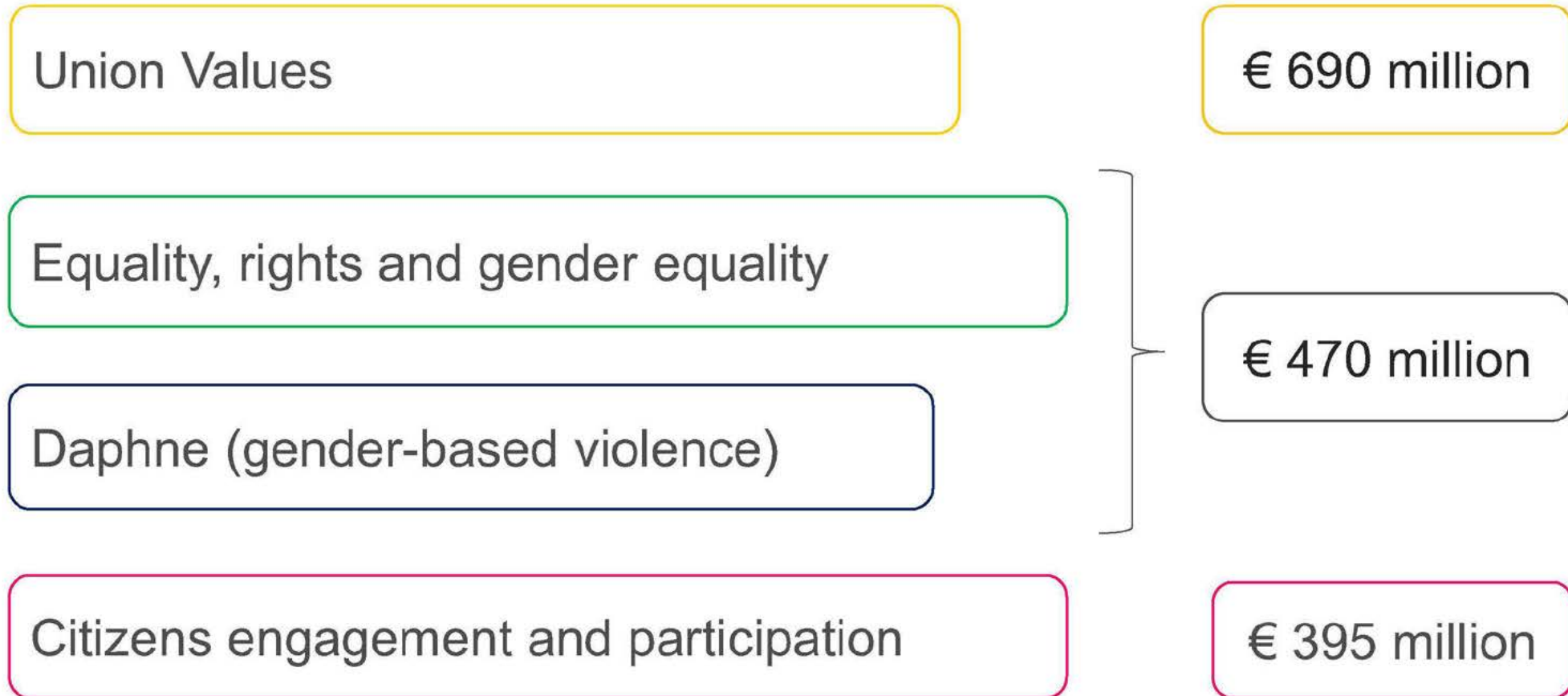
- non-discrimination
 - combat racism
 - rights of the child
 - gender equality
 - rights of persons with disabilities
 - data protection
 - EU citizenship rights
 - fight gender-based violence and violence against children
(former Rights, Equality and Citizenship programme)
- European remembrance
 - Citizens' participation
 - Town-twinning and networks of town
(former Europe for Citizens programme)
 - Union values



1155



Objectives / Strands



Union values

Union Values

Protect, promote and raise awareness on rights by providing financial support to civil society organisations



Equality, rights and gender Equality

Equality, rights and
gender equality



*Gender & non-discrimination
mainstreaming
Gender equality
Non-discrimination
Rights of the child
Rights of persons with disabilities
Union citizenship rights
Protection of personal data*

Citizens' engagement

Citizens engagement and participation



Raise awareness of common history, culture, cultural heritage and values

Promote citizens' participation in and contribution to the democratic and civic life of the Union

Promote exchanges between citizens of different countries (town-twinning & networks of towns)

Daphne

Daphne



Prevent and combat:

- *gender-based violence against women and girls, domestic violence*
- *violence against children, young people, other groups at risk (LGBTQI persons & persons with disabilities...)*

Support and protect all direct and indirect victims of such violence

Which actions?



Training activities, capacity building, mutual learning, workshops, experts' meetings, conferences



Awareness-raising and dissemination activities, media campaigns



Analytical activities, such as studies, researches, surveys and evaluations



Support for European networks, civil society

Participation to the Programme

WHO

- Public and private entities e.g. ministries, towns, civil society organisations
- Legally established in EU Member States (for the moment)
- Profit-oriented and international organisations: limited access (depending also on the call for proposals)

Programme Contact Point



Providing impartial guidance, practical information and assistance to applicants, stakeholders and beneficiaries of the programme

Presentation overview

1. Rights Equality and
- 2. Work Programme 2021-2022**
3. How to apply?
4. Funding & tender opportunities portal
5. Future: Citizens, Equality, Rights and Values

Implementation

a multiannual **work programme 2021-2022** establishing:

- Policy priorities
- Budget
- Activities
- Eligibility, selection and award criteria
- Expected results



Calls for proposals action grants 2021-2022

Protecting and promoting Union values

€ 50.5 million in 2022 - published October 2021

capacity building and awareness on the EU Charter of Fundamental Rights and activities on litigation

€ 2 million in 2022 - published November 2021

Promoting equality and fighting against racism, xenophobia and discrimination

€ 9.9 million in 2021 - deadline 15 June 2021

€ 12.4 million in 2022 - published November 2021

National Roma Contact Points

€ 1.9 million in 2022 - published November 2021



Calls for proposals action grants 2021-2022

Gender equality

€ 6.8 in 2022 million - published November 2021

Rights of the child

€ 2.1 million in 2021 - deadline 7 September 2021

€ 3 million in 2022 - published November 2021

Restricted to Data Protection Authorities on data protection legislation

€ 1.8 million in 2021-2022 - deadline 9 September 2021

Daphne, prevent and combat gender-based violence and violence against children

€ 17.7 in 2021 million - deadline 15 June 2021

€ 27 in 2022 million - published November 2021

Calls for proposals action grants 2021-2022

European remembrance

€ 4.5 million in 2021 - **deadline 22 June 2021**

€ 7 million in 2022 - **published November 2021**

Citizens' engagement

€ 17.5 million in 2022 - **published November 2021**

Town-Twinning and Networks of Towns

€ 6.9 million in 2021 - **deadline 26 August 2021**

€ 11 million in 2022 - **published November 2021**

Union Values

€ 50.5 million in 2022 – apply as of October 2021

Priorities

Supporting, empowering and build the capacity of civil society organisations active in the promotion of values and rights, whose activities help the enforcement of rights under EU law and the EU Charter of Fundamental Rights, including through advocacy and watchdog activities, as well as to promote, safeguard and raise awareness of the Union's values at local, regional, national and transnational level.



Capacity building and awareness on EU Charter and activities on strategic litigation

€ 2 million in 2022 – apply as of November 2021

Priorities

1. Capacity-building and awareness raising on the EU Charter of Fundamental Rights
2. Activities on strategic litigation

Call for proposals to promote equality and to fight against racism, xenophobia and discrimination

CERV-2021-EQUAL: € 9.9 million – deadline 15 June 2021

CERV-2022-EQUAL: € 12.4 million – apply as of November 2021

Priorities

1. Fighting against intolerance, **racism**, xenophobia, discrimination, hate speech and hate crimes
2. Promoting **diversity management** and inclusion at the workplace, both in the public and private sector
3. Fighting **discrimination against LGBTIQ** people and promoting LGBTIQ equality
4. Preventing, reporting and countering **hate speech online**
5. **Restricted to public authorities**

Call for proposals to National Roma Contact Points for Roma equality, inclusion and participation

CERV-2022-NRCP: € 1.9 million – apply as of November 2021

Support National Roma Platforms in:

- providing a forum and encouraging self-mobilisation and cooperation of Roma and pro-Roma civil society organisations;
- address gaps and challenges in the implementation of National Strategies for Roma equality, inclusion and participation;
- generate immediate responses to challenging situations, such as the COVID-19 pandemic.

Applicants must be National Roma Contact Points; only one application per Member State will be accepted; no partners; the EU grant requested cannot be higher than EUR 200 000.



Gender equality

CERV-2022-GE: € 6.8 million in 2022 – apply as of November 2021

Priorities

1. Equal participation and representation of women and men in political and economic decision making
2. Tackling gender stereotypes, including with a focus on youth, and in advertising and the media

Rights of the child

CERV-2021-CHILD: € 2.1 million in 2021 – deadline 7 September 2021

CERV-2022-CHILD: € 3 million in 2022 – apply as of November 2021

Priorities

1. Contribute to the response to the impact of the COVID-19 pandemic
2. Contribute to the implementation of recommendations of the EU Strategy on the rights of the child

Data protection

CERV-2021-DATA: € 1.8 million in 2021 – deadline 9 September 2021

Further support activities carried out by national data protection authorities to facilitate the implementation of GDPR obligations by small and medium-sized enterprises, and to raise awareness about the GDPR among the general public.

Applicants must be national Data Protection Authorities.

European remembrance

CERV-2021-CITIZENS-REM: € 4.5 million in 2021 – deadline 22 June 2021

CERV-2022-CITIZENS-REM: € 7 million in 2022 – apply as of November 2021

Priorities

1. Commemoration of and research and education about crimes committed under totalitarian regimes
2. Resistance, democratic transition and democratic consolidation in the EU

Citizens' engagement

€ 17.4 million in 2022 – apply as of November 2021

Priorities

1. 2024 European Parliament elections
2. COVID-19 crisis and democratic debate/fundamental rights
3. E-democracy
4. Disinformation and other forms of interference in democracy
5. Climate and environment

Town-Twinning and Networks of Towns

CERV-2021-CITIZENS-TOWN-TT & CERV-2021-CITIZENS-TOWN-NT

€ 6.9 million in 2021 – deadline 26 August 2021

€ 11 million in 2022 – apply as of November 2021

I. Town-Twinning

II. Networks of Towns – *Priorities*: EU citizenship rights; tackle discrimination and racism; role of minorities in Europe's cultural development; EU Charter on Fundamental Rights; free, open and healthy democratic governance; free and fair elections in Europe.

Applicants must be towns/municipalities...



Daphne (1)

CERV-2021-DAPHNE: € 17.7 million in 2021 – deadline 15 June 2021

Priorities

1. Early detection, prevention and protection of and/or support to women, children and young people victims or potential victims of violence, with specific attention paid to situations emerging from the COVID-19 pandemic
2. Preventing gender-based violence by addressing masculinities and the engagement of men and boys

Daphne (2)

CERV-2022-DAPHNE: € 27 million in 2022 – apply as of November 2021

Priorities

1. Large-scale and long-term transnational actions
2. Combating and preventing violence linked to harmful practices
3. Primary prevention of gender-based violence
4. National data collection on violence against children
5. Prevent and support child victims of violence

Operating grants

Functioning of a body which has an objective forming part of, and supporting, a Union policy

[CERV-2021-OG-FPA](#) & [CERV-2021-OG-SGA](#)

[CERV-2021-DISA-OG-FPA](#) & [CERV-2021-DISA-OG-SGA](#)

Budget: € 26.7 million in 2021 - deadline 22-29 June 2021

€ 26.7 million in 2022 + € 5.8 million re-granting

Framework Partnership Agreement → operating grant

Presentation overview

Rights Equality and Citizenship

2. Annual Work

3. How to apply?

4. Funding & tender opportunities portal

5. Future: Citizens, Equality, Rights and Values

How to apply



Go to the **Funding & tenders opportunities portal**
[Funding & tenders \(europa.eu\)](http://funding-and-tenders.europa.eu)



Read carefully the **text of the calls for proposals**
e.g. call document [call-fiche_cerv-2021-equal_en.pdf \(europa.eu\)](#)



Questions: EC-CERV-CALLS@ec.europa.eu or EACEA-CERV@ec.europa.eu
For IT related questions: [here](#)
Subscribe to DG Justice newsletter on funding opportunities: [here](#)

Proposal structure

- **Part A** – online form
 - General Information, participants and contacts, budget (90% co-financing & 7% indirect costs)
- **Part B** – pdf document (max. 45 pages)
 - Description of the project and applicant organisation, work packages and activities
- **Part C** – online form
 - Data and indicators
- Annex 1 – CVs key staff applicant and partners (max. 100 pages)
- Annex 2 – Annual Activity Report coordinator
- Annex 3 – List of key previous EU projects of the coordinator for the last 4 years
- Annex 4 – Child protection policy (if contact with children)
- Annex 5 – Letters of support (optional)

Part B

Project summary

- 1. Relevance:** objectives, needs analysis and EU added value
- 2. Quality:** methodology, consortium, staff, management, monitoring and evaluation, cost-effectiveness, risks
- 3. Impact:** impact, communication, sustainability
- 4. Work plan:** work packages and activities, timetable, subcontract
- 5. Other:** Ethics and EU values
- 6. Declarations**

Application forms and annexes

Call data:

Call: **CERV-2021-OG-SGA**
Topic: **CERV-2021-OG-SGA**
Type of action: **CERV-FPA-OG**
Type of MGA: **CERV-SGA-OG**

 Topic and type of action can only be changed by creating a new proposal.

Proposal data:

Acronym: **WPSGA2022**
Draft ID: **SEP-210750316**

Download Part B templates
 Download part B templates

Support & Helodesk

Administrative forms

[Edit forms](#) [Edit Part C](#) [View history](#) [Print preview](#) 

Part B and Annexes

In this section you may upload the technical annex of the proposal (in PDF format only) and any other requested attachments. 

Tpl_Application Form (Part B SEP) (CERV FPA OG)			Upload 
CVs			Upload 
Annual activity reports			Upload 
List of previous projects			Upload 
Child protection policy			Upload 
Other annexes			Upload 

Evaluation

How we evaluate proposals

Evaluation criteria

- Admissibility
- Eligibility
- Exclusion
- Selection
- Award

Admissibility

- Respect deadline for submission of application
- Electronically: SEP Grant Application Form
- Complete: Part A, part B, part C, annexes

Eligibility

- Organisations: legal entity (public or private body)
- Eligible countries: EU Member States
- Project national or transnational
- Applicant with partner(s)
- Minimum requested amount > EUR 75 000
- Project duration max 24 months

Exclusion

Self-declaration in Part A

Applicants/partners cannot participate:

- bankrupt, subject to insolvency or winding-up procedures...
- in breach of their obligations relating to the payment of taxes or social security contributions...
- guilty of grave professional misconduct...
- fraud, corruption, links to criminal organisation, money-laundering...

See Articles 136 and 141 of EU Financial Regulation 2018/1046.



Selection criteria

- **Operational capacity**
 - CVs of the key project staff members – [Annex 1](#)
 - Annual activity report of the applicant – [Annex 2](#)
- **Financial capacity** (no checks for public bodies) → documents NOT to be provided at application stage, but if proposal is selected for funding
 - Balance sheet & loss and profit accounts – 2 years – [Participant register](#)
 - Audit report (grants > 750 000 €) – 2 years – [Participant register](#)
 - Due recovery orders, etc.

Award Criteria

Award criteria	Minimum pass score	Maximum score
Relevance	25	40
Quality	n/a	40
Impact	n/a	20
Overall (pass) scores	70	100

Award criteria (100 points)

a) Relevance (40 / 100; threshold 25 / 40)

- Priorities of the call for proposals
- Needs assessment
- Identification of target groups
- EU added value: implementation of Union law, cross-border cooperation, transfer of good practices and transnational impact
- Synergies and duplication with other Union programmes

! Gender equality and gender & non-discrimination mainstreaming!



Award criteria

b) Quality (40 / 100)

- Methodology = Link between needs - objectives - activities - results; meaningful participation of target groups
- Organisation of work between partners, time schedule
- Risk identification, monitoring & evaluation
- Ethics
- Financial feasibility & cost-effectiveness

! Gender equality and gender & non-discrimination mainstreaming !

Award criteria

c) Impact (20 / 100)

- Results/outcomes, immediate changes on the target groups
- Dissemination strategy
- Multiplier effect
- Sustainability after end of EU funding, long-term impact, long-term socio-economic consequences

! Gender equality and gender & non-discrimination mainstreaming !

What makes a good projects

- ✓ Relevant, in line with policy priorities
- ✓ Addresses a real need in the field
- ✓ Well established methodology
- ✓ Strong partnership (established in advance)
- ✓ Strong EU dimension, EU added value
- ✓ Realistic and sustainable results

Reference documents

<https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents>

- CERV Regulation
- Work Programmes
- Standard application form (in pdf)
- Model grant agreement and contract
- Reporting templates

The screenshot shows the 'Reference Documents' page on the EU Funding & Tenders portal. The top navigation bar includes links for 'SEARCH FUNDING & TENDERS', 'HOW TO PARTICIPATE' (highlighted in yellow), 'PROJECTS & RESULTS', 'WORK AS AN EXPERT', and 'SUPPORT'. Below the navigation bar, the page title is 'Citizens, Equality, Rights and Values Programme (CERV)'. On the left, there is a filter section for 'Programming period' with a dropdown menu showing '2021-2027' and a button to 'Clear filter'. Below this, a blue button labeled 'Citizens, Equality, Rights and Values Programme (CERV)' is visible. The main content area is titled 'Reference Documents' and contains two sections: 'Grants' and 'Procurement'. The 'Grants' section states: 'This page includes reference documents of the programmes managed on the EU Funding & Tenders portal starting with legal agreements and guides for specific actions. Please select the programme to see the reference documents.' The 'Procurement' section states: 'Reference Documents related to tendering opportunities are published on TED eTendering in the calls for tenders.' At the bottom right, there is a search bar with a magnifying glass icon and a 'Filter' button. Below the search bar, a list of document types is displayed, each with a plus icon and a yellow highlight: 'Legislation', 'Work programme & call documents', 'Grant agreements and contracts', 'Guidance', 'Templates & forms', and 'Funding & Tenders Portal'. The footer of the page features the European Union flag and the text 'Commission |'.

List of projects awarded

In each call for proposal webpage → at the bottom

e.g. [Call for proposals to prevent and combat racism, xenophobia, homophobia and other forms of intolerance \(REC-RRAC-RACI-AG-2020\)](#)

Submission service

Get support

Call Updates

Funded project list

Go back to search results

Go to top

Results: 21

Download excel list

Search

TITLE	ACRONYM	PROJECT ID
#NoPlace4Hate: Improving Institutional Response to Hate Speech in Lithuania	LGBT HS	875127
A response to online hate speech through the enhancement of high-quality professional and citizen journalism		
Addressing Modern Antisemitism in the Visegrad countries		
Building a national movement against hate speech that triggers Antisemitism, homophobia, transphobia, xenophobia, and other forms of intolerance		
Building trust and cooperation between law enforcement and LGBTI community to improve response to LGBTI hate crime		
COmmunity work, Participation and Empowerment		
Contra el odio Cotidiano		
Feature A protective environment for LGBTI+ persons		
Hate speech, gender, social networks and political parties		
mutual Learning for Intercultural Appreciation and Strengthened Organisational Networks		

1 2 3 10

Project details

#NoPlace4Hate: Improving Institutional Response to Hate Speech in Lithuania (LGBT_HS)

Project ID: 875127

Programme: REC

Call: REC-RRAC-RACI-AG-2019

D6/Agency: JUST

Abstract

Response by Lithuanian authorities to hate speech is inefficient, because of inconsistent procedural practices, lack of victim support services and low awareness. The proposed project seeks to address this situation through establishing clear guidelines on processing complaints on hate speech within criminal justice system, building capacity of competent public authorities, improving support to victims of hate speech, increasing awareness and improving capacities of two ombudsman-type institutions. The above outlined objectives will be achieved through (1) conducting independent ombudsman review on investigating hate speech, (2) implementing cooperation protocols and conducting internal awareness raising within law enforcement structures, (3) organizing capacity building trainings for law enforcement officers, prosecutors and judges, (4) developing online victim support & information hub (including network of independent service providers and comprehensive referral system), (5) conducting multiple awareness raising campaigns to promote zero tolerance approach towards hate speech. The project will target public authorities in charge of responding to hate speech (1200), victims of hate speech (1000) and general public (1M). It is expected that improved procedures, available victim support services and increased awareness will result both in increased number of reported incidents and reduced prevalence of hate speech. The project will also strengthen institutional capacity of two ombudsman-type institutions to conduct preventive activities in the field of hate speech according to their mandate. The project will produce number of outputs including independent ombudsman review (1), cooperation protocols (2), internal awareness raising campaign for law enforcement (1), capacity building trainings (8), online victim support & information hub (1), public awareness raising campaigns (2), national human rights forum (1).

Activity

No activity available

Keywords:

Projects & results

<https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-results;programCode=REC>

SEARCH FUNDING & TENDERS HOW TO PARTICIPATE **PROJECTS & RESULTS** WORK AS AN EXPERT SUPPORT

Rights, Equality and Citizenship Programme (REC)

Due to a current database failure, you could experience some instability in SyGMA, "My Expert Area", GMS (Grant Management Services) and AMS (Audit Management Services), specially during core business hours. All services have been restored all services by Thursday Morning. We apologise for any inconvenience caused.

Projects & Results

Programming period
2014-2020

Rights, Equality and Citizenship Programme (REC)

Type your keyword...

Project ID
Select a Project ID...

List of funded projects

629 items found

ALLIANCE FOR CHILDREN ON THE MOVE: STANDARD OPERATING PROCEDURES FOR GUARDIANS

Closed
Objectives Implementation of common standards on guardianship systems in four EU countries, safeguarding the rights of unaccompanied and separate

ID	764244	Country	Greece
Acronym	ASOP4G	Contributors	4

Activation and Empowerment of Roma Stakeholders by the National Roma Platform II

Closed
The project consists of Opening and Closing meetings and 3 meetings in regions. The topic of all these meetings will be the identification of the needs and

Thank you



© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.



End

Exhibit 43

CODE OF CONDUCT (+) ON COUNTERING ILLEGAL HATE SPEECH ONLINE

- a) The first Code of conduct on countering illegal hate speech online was adopted on 31 May 2016 by the Commission and Google (YouTube), Facebook, X (formerly Twitter) and Microsoft-hosted consumer services. Between 2018 and 2022, Instagram, Dailymotion, Snap Inc., Jeuxvideo.com, TikTok, LinkedIn, Rakuten Viber and Twitch subscribed to the commitments of the Code (hereafter all the companies subscribing to the Code will be referred to as the “the Signatories”).
- b) Over these years, the Code facilitated key progress, including on the swift review and removal of illegal hate speech content and increased trust and cooperation between the platforms, civil society organisations and Member States authorities in the form of a structured process of mutual learning and exchange of knowledge.
- c) Conscious of the significant opportunities and challenges represented by fast technological development and the evolutions in the policy and legal frameworks in the European Union, the Signatories continue sharing a collective responsibility and pride in promoting the respect of fundamental rights, as enshrined in the EU Charter of Fundamental Rights, in the online environment, as well as the Commission's and EU Member States' commitment to tackle illegal hate speech online.
- d) The Signatories recognise the importance of transparency, monitoring and accountability on their systems and processes as regards the way they deal with alleged illegal hate speech content available on their services, and in particular in relation to their notice-and-action mechanisms. Also, as shown by the success of the collaborative approach under the Code of Conduct since 2016, the Signatories recognise the importance of a multi-stakeholder approach involving representatives of civil society organisations, experts as well as public authorities in the joint efforts to improve content moderation policies, with the aim to assess and mitigate the risk of the public dissemination of illegal hate speech through their services.
- e) The Signatories note that the regulatory framework has advanced with Regulation (EU) 2022/2065 (Digital Services Act, hereafter referred to also as “DSA”)¹ to increase transparency and accountability for intermediary services, informed also by the important achievements of the 2016 Code of Conduct. They recognise the need to further strengthen the Code, building on the DSA, in order to increase its efficiency in fighting illegal hate speech online.
- f) This Code aims to become a voluntary code of conduct under Article 45 of the DSA. It is in this spirit that the Signatories have agreed on this **Code of conduct+** (hereafter referred to also as “the Code”), identifying **voluntary commitments** aimed at creating a framework that facilitates the compliance with and the effective enforcement of the DSA in the specific area of illegal hate speech content, including new measures to address the most recent challenges and threats.
- g) This Code of Conduct is without prejudice to the obligations imposed on signatories under the DSA, ~~as well as the limitations on those obligations provided by the DSA, where applicable.~~

¹ Regulation(EU) 2022/2065 of 19 October 2022 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services

Commented [A1]: Industry: Moved from (h) below as it seems to fit better with this section on scope and purpose of the Code.

Commented [A2R1]: JUST: OK

Formatted: Highlight

Commented [A3]: Industry: Moved up to ensure that we start by stating that the Code is without prejudice to DSA, before covering potential overlaps.

Commented [A4R3]: JUST: OK

Commented [A5]: JUST: As explained in the meeting of 26 Feb, this text will be standard on all CoC, it is a drafting proposal by the Commission legal service. The added sentence does not seem to be needed, it goes without saying that if without prejudice to obligations, that applies also to the limitations of such obligations. We ask for deletion.

The DSA will always take legal precedence for Signatories falling under its scope. ~~This does not affect the voluntary nature of such codes and the freedom of interested parties to decide whether or not to participate.~~

- h) Where commitments overlap with legal obligations that apply to some or all signatories under the DSA, this Code aims to gather specific and proportionate information on how obligations are implemented with respect to illegal hate speech, in particular through the transparency provisions in section 3 below.
- i) Adherence to and compliance with commitments and measures agreed under voluntary codes of conduct by a very large online platform or a very large online search engine where relevant may be considered as appropriate risk mitigation measures under Article 35(1)(h) of the DSA.
- j) To facilitate participation in the Code, Signatories subscribe to implement the relevant commitments through measures that are reasonable, proportionate, and effective in light of the identified risks, the size and nature of their subscribed services, the risk of illegal hate speech on their platforms, the resources available to them and other relevant factors.
- k) The European Commission welcomes the efforts made by the Signatories and will support and monitor the implementation of the **Code of conduct** + in the coming years.

The Signatories of the **Code of conduct**² have agreed on the following public commitments:

1. Terms and conditions for addressing illegal hate speech³

- 1.1 The Signatories are to have in place terms and conditions informing users that they prohibit illegal hate speech on their services. The Signatories are to include in such terms and conditions clear information on the policies around illegal hate speech⁴ on their services, and the measures taken in case of breach of these policies.
- 1.2 The Signatories will inform their users about significant changes made to their terms and conditions that are relevant to how they tackle illegal hate speech, in accordance with Article 14(2) of the DSA.

2 Review and possible removal of or disabling access to illegal hate speech content

² The signatories of the Code of conduct+ are: Facebook, Instagram, YouTube, X (formerly Twitter) TikTok, LinkedIn, Twitch, Viber, Dailymotion, jeuxvideo.com, Snap Inc. and Microsoft-hosted consumer services, as relevant.

³ "Terms and conditions" hereafter refers to terms of service as well as any applicable policies (such as Community Standards, Guidelines, etc). The Signatories acknowledge that there are a variety of valid approaches to crafting Terms and Conditions, e.g. distinctions in how they address illegal content vs. lawful content that violates services' policies, and the Signatories accordingly have discretion in how they comply with commitment 1.1

⁴ 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.

Commented [A6]: JUST: See comment above on the validation of this paragraph by the COM LS. Also, the voluntary nature of the Code is stressed at point h), see further addition highlighted in yellow to further strengthen if needed. Finally, this addition seems to open for the possibility for signatories to adopt the commitments *a la carte*, which has never been the spirit of this Code (even the existing one). "Interested parties" is a terminology that does not appear anywhere else in the text. It seems to repeat the content of Recital 103 of the DSA and not needed here. Suggest a deletion, there appears to be no particular added value.

Commented [A7]: Industry: Reinstated as requested by the EC, using language from EC email 1 March as starting point, with tweaks to ensure clarity.

Commented [A8R7]: JUST: OK

2.1 Pursuant to Articles 16 and 22(1) of the DSA, the Signatories will have in place notice and action mechanisms to allow any user in the EU, including Trusted Flaggers, to notify them of the presence on their service of specific content that the user considers to be illegal hate speech content.

2.2 After receiving and processing a valid notice, the Signatories will review such a notice in a timely, diligent, non-arbitrary and objective manner and, if necessary, act expeditiously to remove or to disable access to the reported content if it is in violation of the Signatories policies, and/or on the basis of the applicable law in accordance with the relevant jurisdiction and in accordance with Article 16 of the DSA.

Commented [A9]: Industry: We have implemented all other edits suggested by the EC in this section, but would ask that 'valid' remains.

Commented [A10R9]: JUST: All the point reads OK

2.3 The Signatories commit to review the majority (at least 50%) of notices received under Articles 16 and 22 of the DSA from Monitoring Reporters within 24 hours. Signatories will apply their best efforts to go beyond this target and aim for two-third (at least 67%) of those notices subject to Annex 1. Signatories are free to set up the processes they see fit to reach this target, such as designated reporting processes. This acknowledges that the time it takes to action notice may depend on contextual factors, such as the volume, severity, virality, and complexity of the content flagged.

Commented [A11]: Industry: Aligned with Annex 1 as discussed.

3 Transparency, accountability and monitoring.

3.1 The Signatories agree to assess the adherence to the specific commitments on the review of notices under 2.3 and to monitor trends over time, following a methodology agreed between the Signatories and the European Commission (See Annex 1).

3.2 Beyond the monitoring on the review of notices under 2.3, the Signatories commit to provide additional information accompanying their results of each monitoring exercise, as referred to in Annex 2.

4 Intra-industry and multi-stakeholder cooperation to address and prevent risks of the spread of illegal hate speech

4.1 The Signatories will continue to implement and strengthen, where relevant, their partnerships with Monitoring Reporters, as defined in Annex 1.

4.2 The Signatories will participate in regular exchange fora aimed at facilitating an effective flow of best practices, expertise, and tools on measures to address the dissemination of illegal hate speech content through their services. These fora will be made up of:

4.2.1 An annual convening of Signatories and Monitoring Reporters organised by the European Commission with input from the Signatories.

4.2.2 An exchange of best practices among the Signatories, materialised in a separate session during the annual convening in 4.2.1.

4.2.3 A structured dialogue among Signatories and Monitoring Reporters to exchange information on trends and developments related to illegal hate speech online. This may be aided by a set of standard questions to answer when sharing a trend with the Signatories.

Commented [A12]: Industry: Upon re-read, to ensure this is not confused with escalation channels/reporting processes.

4.2.4 A shared online knowledge hub facilitated and populated by the European Commission and with adequate safeguards to protect information, gathering jurisprudence and legal resources useful to assess the legality of alleged illegal hate speech, and/or past reports under the Code. Signatories or the Monitoring Reporters may contribute through and best

Commented [A13]: Industry: As discussed in 26 Feb meeting.

practice resources from the Signatories and Monitoring Reporters, including relevant information on the challenges and opportunities presented by emerging technologies in both the generation and detection of illegal hate speech.

4.3 In the exceptional event of extraordinary circumstances leading to a serious threat to public security or public health in the Union or in significant parts of it with an imminent and material risk of a significant increase in the dissemination of illegal hate speech on the Signatories' platforms, the Signatories and the European Commission can initiate a meeting under the umbrella of the exchange fora per 4.2 (or more, should it be necessary and proportionate) to discuss trends, exchange information, and explore avenues of cooperation.

4.3.1 Signatories' commitment to participate participation will depending on their level of exposure to the imminent and material risk of significant increase in the dissemination of illegal hate speech on their platform. The Signatories and the European Commission can decide together whether to involve Monitoring Reporters in those discussions.

4.3.2 All efforts shall be taken to ensure these meetings do not duplicate or otherwise overlap with legal requirements (such as Art 36 DSA), voluntary processes (such as the EU Code of Practice on Disinformation) or other requests. Meetings as described in 4.3 will might not be initiated if some the Signatories are already engaged in similar meetings on the same subject matter with the Commission under Art. 36 DSA.

4.4 Signatories commit, including through existing initiatives, to exchange between their relevant teams, where appropriate, about cross-platform operations and relevant trends that emerge on their respective services, with the aim of preventing the dissemination of illegal hate speech on other services, in full compliance with privacy legislation and with due consideration of security threats and fundamental rights threats.

5 Awareness raising, online civility and counter narrative initiatives

5.1 Recognising the value of raising awareness against hateful rhetoric and prejudice and to foster online safety and civility, the Signatories will continue supporting tools and approaches on counter and alternative narratives, new ideas and initiatives as well as educational programmes that encourage civility online and critical thinking. These will complement initiatives funded and facilitated by the European Commission, such as in the context of the Citizens, Equality, Rights and Values Programme and the High Level Group on combating hate speech and hate crime

5.2 Signatories will make best efforts to provide information about training, resources and support available to Trusted Flaggers and civil society organisations regarding best practices and how to use online platforms for delivering educational and awareness raising campaigns.

Commented [A14]: JUST: The Commission, as discussed, will set up the hub but we prefer a more general formulation as third parties (e.g. a contractor) may be asked to find and "populate" the content with our validation. If there is a need to clarify the exact role of signatories in the context of the knowledge hub, we propose the clarification further below to be clear on what contribution is expected.

Commented [A15]: Industry: Re-drafted on the basis of the discussion on 26 Feb and taking into account the suggestions shared by the EC in 1 March 2024 email.

Formatted: Highlight

Commented [A16]: JUST: This Code cannot establish a "mechanic" association between extraordinary circumstances and the imminent risks. It seems to interpret Article 36 in a restrictive way. As discussed, the added value of a meeting called under 4.3 is to be complementary, yet not duplicate (as stated in current 4.3.2) crisis mechanism under Art 36. We feel the current version after tracks is sufficiently clear in saying a) this could be extra exceptional meeting(s) (see addition in yellow), nothing more demanding than that; b) meetings might not be initiated if some signatories are already engaged in meetings on the same subject matter under Art. 36. If we all read it the same way, can this compromise text be accepted?

Commented [A17]: JUST: This duplicates the meaning of the following sentence in relation to Art 36 DSA. It also reads too restrictive, as it precludes the possibility (unfortunately quite likely in times of instability such as the ones we are leaving in) that parallel, yet very diverse, exceptional events with imminent risks could lead to the need of meetings under both the Disinfo Code and this one (e.g. big interference in elections + a war that causes spikes of hatred at the same time). We suggest to delete and offer a compromise text for the remaining sentence to mirror the above.

Exhibit 44

Question and answers on the Code revision

27 March 2023

- How do you envisage the Code working alongside the risk assessment requirements? **The Code itself, and therefore the text martyr initially drafted in particular the section on “Cooperation between IT Companies, trusted flaggers and experts to address and prevent risks of the spread of illegal hate speech”, intend to play a part as a mitigation measure under the DSA risk assessment requirements. We would welcome additional suggestions on how this could be further elaborated either in the Code commitments or in the initial statement / preamble.**
- Does the Commission envision a formal connection to DSA Recital 90 and the collaborative dialogue with trusted flaggers and experts? **The initial statement / preamble may refer to recital 90 of the DSA. Please do not hesitate to suggest draft initial wording for that part too (currently blank).**
- Will the industry signatories be the same as those already signed up to the existing Code? **The negotiation phase on the new Code will be conducted with the industry signatories that have signed up to the existing Code. This will not prevent other signatories to join once the new Code is agreed. In fact, we would expect a number of new platforms to join later**
- Will the trusted flaggers need to revise their status and also be trusted flaggers under the DSA? **Not necessarily, the Code prompts the existence of trusted flaggers programmes set up by platforms which can be complementary to or overlap with the DSA trusted flaggers. See proposed commitments under the section on “Cooperation between IT Companies, trusted flaggers...”**
- Do you agree that the Code should continue to only cover illegal hate speech in line with its focus to date? **Yes, we agree.**
- Do you agree that the Code should make clear that outcomes related to the monitoring exercise do not necessarily imply systemic failure to meet DSA obligations, particularly in light of the fact that notices sent during monitoring rounds may relate to gray-area content, and violate neither local/EU law nor the company’s Terms and Conditions? **The monitoring exercises are only meant to assess the performance on the Code commitments related to notice-and-action. A revised code of conduct could also better reflect the discussion and questions on grey area content or other factors that may impact on removal rate or time of assessment. Grateful for any suggestions you may have that would allow for the monitoring and information from the platforms to show a more granular or complex picture of the situation**
- Do you agree that any researchers seeking access to data under the revised Code will need to follow the requirements and procedures in DSA Art. 40 for researcher access to data? **Yes we agree.**
- Do you agree that standardized KPIs should match the DSA’s transparency provisions, e.g., in Art. 15 and Art. 24? **Yes we agree and you are invited to suggest wording in text martyr in this direction, including the initial statement / preamble**
- Do you intend for the Code to be negotiated between existing platform signatories and the Commission and that the Commission will consult the HLG on Hate Speech at a later date once the revised Code is near final? This is how we’re interpreting the draft Rules of Engagement. **Yes, this is how we envisage the process.**

Exhibit 45

CODE OF CONDUCT (+) ON COUNTERING ILLEGAL HATE SPEECH ONLINE

Commented [A1]: Provisional name "Code of conduct +"

- a) The first Code of conduct on countering illegal hate speech online was adopted on 31 May 2016 by the Commission and Google (YouTube), Facebook, Twitter and Microsoft-hosted consumer services. Between 2018 and 2022, Instagram, Dailymotion, Snap Inc., Jeuxvideo.com, TikTok, LinkedIn, Rakuten Viber and Twitch subscribed to the commitments of the Code (hereafter all the companies subscribing to the Code will be referred to as the "the Signatories").
- b) Over these years, the Code facilitated key progress, including on the swift review and removal of illegal hate speech content and increased trust and cooperation between the platforms, civil society organisations and Member States authorities in the form of a structured process of mutual learning and exchange of knowledge.
- c) Conscious of the significant opportunities and challenges represented by fast technological development and the evolutions in the policy and legal frameworks in the European Union, the Signatories continue sharing a collective responsibility and pride in promoting the respect of fundamental rights online, as well as the Commission's and EU Member States' commitment to tackle illegal hate speech online.
- d) The Signatories recognise the importance of transparency, monitoring and accountability on their systems and processes as regards the way they deal with alleged illegal hate speech content available on their services, and in particular in relation to their notice-and-action mechanisms. Also, as shown by the success of the collaborative approach under the Code of Conduct since 2016, the Signatories recognise the importance of a multi-stakeholder approach involving representatives of civil society organisations, experts as well as public authorities in the joint efforts to improve content moderation policies, assess the risks represented by illegal hate speech and to design appropriate mitigation measures in respect of the spread of illegal hate speech on their services.
- e) The Signatories note that the regulatory framework with the Digital Services Act¹ (hereafter referred to also as "DSA") has advanced to further transparency and accountability for online platforms also informed by the the important achievements of the 2016 Code of Conduct and are building on these, and they recognise while recognising the need to further strengthen the Code, building on the DSA, in order to increase its efficiency in fighting illegal hate speech online.
- f) It is in this spirit that the Signatories have agreed on this **Code of conduct+** (hereafter referred to also as "the Code"), identifying additional commitment areas and new measures to address the most recent challenges and threats and with the Digital Services Act (DSA) as the new regulatory framework in place.
- g) ~~Actions under the Code will complement and be fully aligned with regulatory requirements and overall objectives in the DSA. The DSA sets out a regulatory framework for content moderation that is harmonised across the EU, including issues related to illegal hate speech.~~
- h) ~~This Code aims to become a voluntary code of conduct under Article 45 of the DSA. Where relevant, compliance with its commitments and measures under this Code may serve as risk mitigation measures for Very Large Online Platforms to address relevant systemic risks identified under Article 34 of the DSA.~~

¹ Regulation (EU) 2022/2065 of 19 October 2022

1.1 To facilitate participation in the Code, Signatories are able to subscribe to the commitments and to implement them through measures that are proportionate in light of the size and nature of their services, the risk of illegal hate speech on their platforms, and the resources available to them.

1.2 The European Commission welcomes the efforts made by the Signatories and will support and monitor the implementation of the **Code of conduct** + in the coming years.

The signatories of the **Code of conduct**+² have agreed with the European Commission on the following public commitments:

1. Terms and conditions for addressing illegal hate speech³

1.1 The Signatories are to have in place terms and conditions informing users that they prohibit illegal hate speech on their services. The Signatories are to include in such terms and conditions clear information on the policies around illegal hate speech⁴ on their services, and the measures taken in case of breach of these policies.

1.2 In accordance with Article 14 of the Digital Services Act (Regulation (EU) 2022/2065 of 19 October 2022, hereafter the "DSA"), the Signatories will ensure that their terms and conditions are set out in clear, plain, intelligible and user-friendly language, and shall be publicly available to users of the service. The Signatories shall also pay due regard to the fundamental rights and freedoms as enshrined in the EU Charter of Fundamental Rights, when applying and enforcing these terms and conditions.

1.2.1 When applying and enforcing the restrictions outlined in the terms and conditions in accordance with the DSA, the Signatories will base themselves on applicable law, 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.

Formatted: Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Commented [A2]: It can be simplified to avoid unnecessary repetition of DSA. See footnote added as we would need to maintain a reference to the FD.

2 Review and possible removal of or disabling access to illegal hate speech content

2.1 Pursuant to Article 16 of the DSA, the Signatories will have in place mechanisms to allow any user in the EU to notify them of the presence on their service of specific content that the user considers to be illegal hate speech content.

² The signatories of the Code of conduct+ are: Facebook, Instagram, YouTube, TikTok, LinkedIn, Twitch, Viber, Dailymotion, jeuxvideo.com, Snap Inc. and Microsoft-hosted consumer services, as relevant.

³ "Terms and conditions" hereafter refers to terms of service as well as any applicable policies (such as Community Standards, Guidelines, etc). Acknowledging platforms' terms and conditions can be global in nature, and without prejudice for the distinction under the DSA between platform's obligations in relation to illegal content on the one hand and harmful but legal harmful content on the other.

⁴ 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.

2.2 After receiving and processing a valid notice in accordance with Article 16 of the DSA, the Signatories will review such notice in a timely manner and, if necessary, act expeditiously to remove or to restrict access to the reported content if it is in violation of the Signatories policies, and/or on the basis of the applicable law in accordance with relevant jurisdiction and in accordance with Article 6 of the DSA.

2.3 The Signatories will commit to reviewing in less than 24 hours two third of the sufficiently precise and adequately substantiated notices on content alleged to be illegal hate speech in less than 24 hours, in the majority of cases⁵. Other types of illegal content may require longer or shorter timelines for processing of notices, which will depend on the facts, circumstances and types of illegal content at hand.

Commented [A3]: We think we can be more ambitious in the CoC+, looking also at the performance on this particular commitment over the last years, e.g. data from monitoring exercises.

Commented [A4]: It may be confusing to refer to other types of illegal content in a Code focused on hate speech.

3 Transparency, accountability and monitoring,

Formatted: Font color: Auto

3

Formatted: Indent: Left: 0.25", No bullets or numbering

3.1 The Signatories shall will provide transparency reports in accordance with Articles 15, 24 and, where applicable, 42 of the DSA as well as any implementing act adopted pursuant these Articles.

3.2 The Signatories and the European Commission agree to assess the adherence to the specific commitments on the review of notices under 2.3 and to monitor trends over time, following a methodology agreed between the with Signatories and the European Commission (See Annex XX).

Commented [A5]: The EC agrees on the methodology for assessing, rephrased to fine tune the meaning.

3.3 Beyond the monitoring on the review of notices under 2.3, for the purpose of the monitoring exercise and the transparency reporting under the DSA referred to in 3.1, the Signatories will commit to provide ~~already best efforts to provide~~ additional meaningful and comprehensible information on the measures taken to address illegal hate speech as a result of their content moderation policies based on a structured set of guiding questions (Annex YY).

Commented [A6]: Code text indeed needs to be ambitious and "best efforts" sounds slightly weaker. The Annex YY will define what additional information will be.

4 Intra-industry and multi-stakeholder cooperation to address and prevent risks of the spread of illegal hate speech

4.1 The Signatories will continue to implement and strengthen, where relevant, their partnerships with civil society organisations, including those which may act as trusted reporters under separate voluntary programs outside of DSA (hereafter "civil society organisations"), in addition to the role of and engagement with the trusted flaggers, whose status has been awarded according to the provisions in Article 22 of the DSA for their expertise in illegal hate speech (hereafter "Trusted Flaggers").

4.2 The Signatories will participate in regular exchange fora aimed at facilitating an effective flow of best practices, expertise, and tools on measures to address illegal hate speech aligned with the DSA. These fora can support both the Signatories' mitigation of the systemic risks⁶ of disseminating illegal hate speech if and where relevant, and the balancing exercise the Signatories must undertake with relevant fundamental rights as enshrined in the EU Charter of fundamental rights. These fora will be made up of:

4.2.1 An annual convening of signatories, Trusted Flaggers and other pertinent civil society organisations, organised by the European Commission with input from the Signatories.

⁵ Reference is made to recital 6872 in the DSA which refers to the 2016 Code of conduct on countering illegal hate speech online as setting a clear albeit not legally binding benchmark for the time needed for the review of notices on illegal hate speech content.

⁶ To take effect if and when the Code formally becomes a Code under the DSA.

- 4.2.2 An exchange of best practices among the Signatories, materialised in a separate session during the annual convening in 4.2.1.
- 4.2.3 A structured channel among Signatories and Trusted Flaggers, relevant civil society organisations under Article 4.1 above, to exchange information on trends and developments related to illegal hate speech online. This may be aided by a set of standard questions to answer when sharing a trend with the Signatories.
- 4.2.4 A shared online knowledge hub facilitated by the European Commission and with adequate safeguards to protect information, gathering past reports under the Code and best practices resources from the Signatories, Trusted Flaggers, other pertinent civil society organisations, and relevant experts.
- 4.3 Signatories ~~shall apply best efforts~~commit, including through existing initiatives, to exchange between their relevant teams, where appropriate, about cross-platform operations, and relevant trends that emerge on their respective services, with the aim of preventing dissemination on other services, in full compliance with privacy legislation and with due consideration for security threats and fundamental rights threats.

Commented [A7]: See previous comment on need to be ambitious.

5 Awareness raising, online civility and counter narrative initiatives

- 5.1 Recognising the value of raising awareness against hateful rhetoric and prejudice and to foster online safety and civility, the Signatories and the European Commission will continue supporting tools and approaches on counter and alternative narratives, new ideas and initiatives as well as educational programmes that encourage civility online and critical thinking.
- 5.2 The Signatories will inform their users about significant changes made to their Terms and Conditions, in accordance with article 14 of the DSA.
- 5.3 Signatories will make best efforts to provide information about training, resources and support available to Trusted Flaggers and civil society organisations on best practices and how to use online platforms for delivering educational and awareness raising campaigns.