ALGORITHMIC AMPLIFICATION

REDUCING RADICALIZING ONLINE CONTENT IN EUROPE

PREPARED BY: FARA ISLAM, MPP

APRIL 2022

Trust Lab | FRANK BATTEN SCHOOL *of* LEADERSHIP *and* PUBLIC POLICY

CONTENTS

Fara Islam

Master of Public Policy Candidate

Frank Batten School of Public Policy and Leadership

University of Virginia

Prepared for Trust Lab, Inc.

Client Overview

Trust Lab is a startup company based in Palo Alto, California that specializes in products in the world's largest social media platforms, online marketplaces, and apps to protect users against misinformation, hate speech, identity fraud, and other harmful content. Trust Lab uses ML-based classifiers and rules engines built by Trust & Safety experts from Google, YouTube, Reddit and TikTok to identify and measure high-risk and harmful content, accounts, and transactions at scale. The majority of the leading social media companies use Trust Lab's innovative tools and services, as do leading marketplaces, messaging companies, as well as content hosting services of different sites. Trust Lab also works with both the U.S. Government and the European Union to protect free speech online while suppressing harmful content. Trust Lab was chosen by the European Commission to track and measure extremist and violent content across European Union Member States.

Mandatory Disclaimer

The author conducted this study as part of the program of professional education at the Frank Batten School of Leadership and Public Policy, University of Virginia. This paper is submitted in partial fulfillment of the course requirements for the Master of Public Policy degree. The judgments and conclusions are solely those of the author, and are not necessarily APP II Assignments APP Technical Report endorsed by the Batten School, by the University of Virginia, or by any other agency.

Acknowledgments

To The Most High, nothing is possible without Your Guidance. Alhumdulilah.

To my parents, Shahin and Amar Islam, thank you for your constant love, support, encouragement, and sacrifice.

To my sister, Urmana, thanks for always pushing me to reach beyond the stars.

Executive Summary

From 2010 to 2021, there have been 1,871 terrorist attacks in the European Union (Statista, 2023), with a total of 524 preventable deaths. Social media has become a primary distribution channel for terrorism and violent extremism (TVE) content. In fact, a survey with European adolescents demonstrates that nearly 80% of the participants have been confronted with content inciting TVE at least once (Bécuwe et al., 2018). This is an issue manifested online in social media platforms, which has created escalation to real world harm in Europe. Recommender algorithms provide a personified feed to the user dependent on their interests. A personalized feed can develop the more a user searches on a particular topic, interacts with its content, and actively seeks it. When an individual consistently and actively looks for harmful content, they have the risk of being connected with like-minded individuals that can amplify these ideologies through filter bubbles. Automated recommender algorithms from social media platforms pose the potential risk of radicalizing users when it comes to terrorism and violent extremism (TVE) content in Europe.

I propose three alternatives to address this problem:

1. The Status Quo
2. Establishing baseline definitions for TVE and Borderline Content
3. Algorithmic Modification

These alternatives are evaluated through by the follow criteria:

1. Effectiveness (30%)
2. Political Feasibility (30%)
3. Corporate Feasibility (20%)
4. Transparency (20%)

Based on this assessment, I conclude that algorithmic modification is the best way to stop violent extremism and terrorism content from spreading on the internet, as well as radicalizing users. This approach scores highly favorable on effectiveness and political feasibility, then favorable on corporate feasibility and transparency.

Acronyms

DGA	Data Governance Act
DMA	Digital Markets Act
DSA	Digital Services Act
EC	European Commission
EU	European Union
GDPR	General Data Protection Regulation
RS	Recommender System
TVE	Terrorism and Violent Extremism

Definitions

Borderline: Content that comes close to violating policies around terrorism and violent extremism and that shares some characteristics of hateful or harmful content (Thorley et al., 2022).

Echo Chambers: Refer to groups of individuals who share similar beliefs, and these groups can often become polarized, limiting exposure to different perspectives and hindering the ability to receive counter-messaging. As a result, individuals may become entrenched in their own radical ideologies (Wolfowicz, 2021).

Filter Bubbles: Algorithms determine a user's future online experiences by selecting and prioritizing certain content based on their perceived preferences. This can result in the promotion or recommendation of specific content to the user (Wolfowicz, 2021).

Radicalization: Individuals can adopt beliefs that not only justify violence but also compel them to take violent actions through their interactions with and exposure to various types of internet content and groups. This process involves being influenced by these online sources to the extent that individuals feel compelled to act on their beliefs (UNODC, 2018).

Introduction

In 2021, the number of terrorist attacks in Europe was at a record low of 15 attacks (Statista, 2023). In 2017, the European Union strengthened control of the acquisition and possession of firearms; thus explaining the steep decline in terrorist attacks (Council of the European Union, 2017). However, the constant usage of social media has become a primary distribution channel for terrorist and violent extremist content. While gun regulation has reduced terrorist attacks over time, the same cannot be determined for radicalization overall (Council of the European Union, 2022). The use of automated recommender algorithms on social media platforms is particularly concerning, as they increase the risk of radicalizing users with TVE content in Europe.

The internet has brought extensive change in peoples' lives. It has revolutionized how we communicate and simplified the way we create networks among like-minded individuals. Recommender algorithms from social media companies provide a personified feed to the user dependent on their interests. A personalized feed can develop the more a user searches on a particular topic, interacts with its content, and actively seeks it (Meserole, 2022). The personalization of topics can vary from cooking, sports, but even harmful rhetoric. When an individual is constantly faced with their interests, like extreme violence, they have the risk of being connected with like-minded individuals that can amplify these ideologies (Doxsee et al., 2022).

The main concern is that users who become radicalized towards terrorism and violent extremism do so because of their own motivation (Doxsee et al., 2022). These users actively search for harmful content instead of being influenced by the content that is randomly recommended to them on their social media feeds. However, the terrorist and extremist content they search for is amplified from their consistent interactions, as shown in Figure 1.

Figure 1: 10 Steps to Extremism

Source: Fara Islam, 2023.

From stakeholder interviews with behavioral and AI experts, this 10 step path to extremism was created to showcase how an individual can manifest radicalization through the internet. This was made to emphasize that gun regulation laws in the EU were able to stop radicalization in Step 8, acquire weaponry, from occurring at an alarming rate; however, Steps 1- 7 still occur in a way that allows users to create harmful ideologies through the platforms they use.

The aim of this report is to help understand a dangerous problem between social media companies, regulations, and its constituents to suggest policies to address safety measures. It will start by defining the problem and providing background information, including existing legislation and the usage of algorithms. Then, it will examine academic literature to highlight potential solutions. Finally, it will evaluate each policy option to determine the most effective solution for the future.

Problem Statement

The European Union experienced a total of 1,871 terrorist attacks from 2010 to 2021, which resulted in 524 preventable deaths (Statista, 2023). However, the number of terrorist attacks decreased significantly in 2021, with only 15 incidents reported. The decrease can be attributed to the European Union's strengthened control over the acquisition and possession of firearms in 2017 (European Council, 2017). While gun regulation has contributed to reducing terrorist attacks, it does not prevent the root of these radicalized thoughts from occurring (Council of the European Union, 2022). Unfortunately, social media continues to be a significant platform for the distribution of terrorist and violent extremist (TVE) content. In fact, a survey conducted with European adolescents found that nearly 80% of the participants had been exposed to such content at least once (Bécuwe et al., 2018). **Automated recommender algorithms from social media platforms pose the risk of radicalizing users when it comes to terrorism and violent extremism content in Europe.**

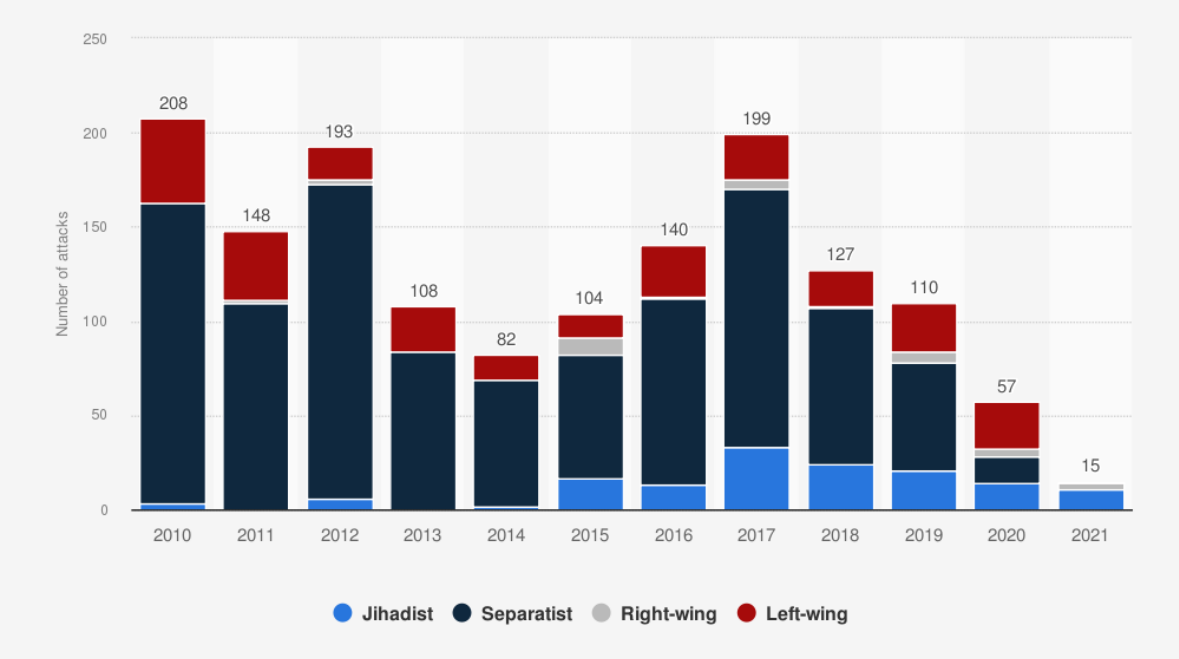


Figure 2: Number of terrorist attacks in the European Union (EU) by affiliation, 2010 - 2021

Source: Statista, 2023

Background

Extremism in the European Union

The deadliest terrorist attack in the European Union was the Lockerbie Bombing on December 21, 1988 in Scotland. After a bomb was detonated, all 259 people on board were killed, and 11 individuals on the ground also died (Statista, 2023). It was believed by investigators that the bombing had been carried out by two Libyan intelligence agents, and there was much speculation that the attack was in retaliation for the U.S. bombing campaign against Tripoli, which took place in 1986 (Britannica, 2023). A more recent deadly attack in the Western world was the Christchurch shooting on March 15, 2019 in New Zealand. In the span of 36 minutes, an Australian far-right extremist had fatally shot 51 Muslims in two mosques in Christchurch, with plans to target a third mosque (BBC, 2020). The horrific event was live streamed by the terrorist for 17 minutes, and was viewed over 4,000 times before being removed (Flynn, 2019), which can be traumatizing or encouraging based on the user watching.

The presence of extreme polarity is what is igniting extremist acts in the EU. In terms of platform influence, it can be recognized that the internet connects millions of individuals and opens opportunities to radicalize them with the existence of channels that reinforce extreme ideologies. Due to confirmation bias, individuals will constantly expose themselves to like-minded individuals as a form of validity (Doxsee et al., 2022). This phenomenon can fortify claims and sentiment so intensely, that its widespread disinformation at an escalating rate that can't be rectified properly since it is not recognized as disinformation (Meserole, 2018). Through filter bubbles, they form communities with similar extremists as themselves. A filter bubble is when an algorithm selectively limits the information that a user is exposed to based on their interests and past interactions on the platform (Pariser, 2011). This is extremely present with the wake of social media platforms, and in essence, tends to normalize behaviors and attitudes that otherwise may carry a risk of being considered unacceptable or inappropriate in the physical world.

Current European Union Jurisprudence

Until the Digital Services Act (DSA) is enacted in 2024, there are no active European Union acts or laws that directly address the reduction of TVE content by filter bubbles or algorithmic amplification. However, active regulations like the General Data Protection Regulation (GDPR), Digital Markets Act (DMA), and Data Governance Act (DGA) are regulations that aim to protect the privacy and data of individuals and regulate online platforms and tech companies (Paeman, 2022). While these regulations do not have a direct relationship with terrorism, they may indirectly affect counter-terrorism efforts in certain ways. For example, under the GDPR, companies are required to obtain explicit consent from individuals before collecting or processing their personal data, which includes data related to potential terrorist activities. This requirement may make it more challenging for law enforcement and intelligence agencies to access such data in investigations related to terrorism (ICO, 2018). On the other hand, the DMA and DGA seek to increase transparency and fairness in online business transactions, which may indirectly help prevent enabling or facilitating the transfer of funds to terrorist organizations.

The DSA is the first EU act that includes provisions that seek to improve transparency and accountability in the way online platforms personalize and distribute content. This is directly related to how social media algorithms are made to form filter bubbles through content. The act requires companies to provide transparency reports to the European Commission about the algorithms they use to personalize content. The DSA is different compared to the acts listed above because it is centered on regulating the content that platforms distribute rather than how they collect and use consumer data (Mackrael, 2022).

The DSA enhances previous regulatory efforts by making social media companies accountable for the problems they have caused. New Zealand enacted the Christchurch Call to Action Initiative that requires tech companies to remove harmful content like extremist videos and images quickly or face fines (Canlas, 2019); this action was in response to the Christchurch shooting in 2019. The DSA echoes similar practices of the New Zealand act by increasing transparency in content moderation and addressing various types of harmful content (Morar, 2022). However, the DSA's approach to content

moderation does not consider the impact of influential individuals within social media networks. This means that an influencer with a large following could share a post containing hate speech or TVE content, which could then be shared and amplified by their followers, even if the platform's algorithm limits the promotion of such content. (Malone, 2022). Additional measures would be needed to address the influence of these individuals on social media networks.

Technical Understanding of Recommender Systems

Every social media platform possesses an algorithm. Some algorithms can prioritize showing trending content on your feed such as Twitter, while others fill your feed with posts made from who you follow. A recommender system (RS) is a specific type of algorithm that is used to make personalized recommendations to users based on their past behavior, preferences, and interactions. RSs are categorized into six types: collaborative filtering, content-based, utility-based, demographic-based, knowledge-based, and hybrid-based. Among these, the most traditionally used approaches are content-based and collaborative filtering.

Recommender System Process

The recommendation system determines what set of content populates to a user's feed based on the interaction they make with certain content, this is done by a rating system of what the user is 'interested' in. Recommender systems (RSs) are designed to evaluate the rating function fR for a specific user U and item I. The technical formula is understood as: $fR: U \times I \rightarrow R$. This function takes into account the user's preferences and the characteristics of the item to generate a rating that reflects how well the item matches the user's interests. The rating function fR is a mapping from the set of all possible user-item pairs ($U \times I$) to a real number rating for R, which represents the rating or score assigned to the item by the user (Suhaim & Berri, 2021).

Traditional recommender systems, such as collaborative filtering, content-based filtering, and hybrid filtering, rely on the two-dimensional relationship between users (U) and items (I) to generate recommendations. Collaborative filtering algorithms look for patterns in the ratings and behavior of similar users to make recommendations, while content-based filtering algorithms use the features of the

items to recommend similar items to users. Social networks provide additional information that traditional recommender systems don't consider, such as friends and followers of users, which can improve the quality of personalized recommendations. The homophily principle assumes that people who are friends in social networks have something in common, creating correlations between users (Fayyaz et al., 2020). Personalized recommendation is based on the analysis between users and items to reflect varying interests.

Collaborative-Filtering Recommendation Systems

Collaborative filtering is a way that websites like Amazon, Netflix or YouTube recommend products or content to users based on their past behaviors or ratings. This method creates a database of user preferences and tries to find similar users who have rated similar products or content. By doing this, the website can generate recommendations that the user may be interested in based on the preferences of similar users. Collaborative filtering can be split into two types: item-based filtering and user-based filtering (Fayyaz et al., 2020). User-based filtering is the most common and involves finding similar users to the target user to create recommendations (Eirinaki et al., 2018). One of the benefits of collaborative filtering is that it does not require knowledge about the items being recommended, it just focuses on user behavior and preferences.

Content-Based Recommendation Systems

Content-based filtering is another way that websites like Amazon or Netflix recommend products or content to users based on their past behaviors or ratings, but instead of looking for similar users, it looks at the user's own preferences and builds a profile based on keywords and tags. This method measures the similarity between the user profile and items to predict ratings on new or unseen items (Zhang, Lu & Jin, 2021). One of the benefits of content-based filtering is that it does not require data from other users, so it's scalable and can handle many users. Content-based filtering needs a lot of knowledge about the items being recommended, and it may not be accurate if there's not enough information to differentiate between similar products. Additionally, it may not offer much variety in recommendations as it tries to match the features of the user's profile and items.

Consequences of the Problem

Terrorism has been a growing concern for governments worldwide, and the increased use of social media platforms as a tool for radicalization and recruitment has made it even more challenging to combat. With the rise of global communication and the internet's borderless nature, the effects of social media terrorism can be felt across continents, including Europe. The spread of extremist ideologies, the recruitment of vulnerable individuals, and the promotion of violent acts can all occur through social media platforms and have a significant impact on European countries' safety and security. The first example below relates the author's personal connection to gun violence and violent extremism with a recent tragedy at the University of Virginia; while the second example relates to a terrorist attack in Paris, France.

Mass Shooting in UVa on November 13, 2022

In most recent news, Twitter has derailed from the traditional limitations and policies of online hate speech and extremism since Elon Musk has announced his full ownership over the social media giant (Dang, 2022). In efforts to dismantle the limitations on free speech, Musk has opened the gates of unfiltered and extremist sentiments. Within the first hours of Musk's stewardship of Twitter have been dominated by his supporters relishing their ability to use profane slurs, racial epithets, and a torrent of racist, antisemitic, Islamophobic, homophobic, and transphobic hate speech (Gilbert, 2022; Patel, 2022). Figure 3 examines an example of how Elon Musk's ownership of Twitter has created a sense of chaos on the platform where it's harder to detect true information. This is dangerous because of the audience and community it attracts. Given the real-time nature of this, there is no reactive real-world outcome that can correlate with an attack of extremism; however, the online presence of such speech can encourage real-world sentimental actions of it.

Figure 3: Screenshot from the hours of lockdown during UVa's November 13, 2022 mass shooting. Not only can this encourage individuals to participate in extremist acts as they witness the streaming, but traumatized individuals as they experience the event in real-time.

Google vs. Gonzalez

Recommender systems can create filter bubbles and echo chambers that can limit their exposure to diverse perspectives and opinions, leading to a lack of critical thinking and a narrowed worldview. Not only can this limit their perspective, but can manifest hate and resentment towards those with differing beliefs or vulnerable groups depending on their worldview. As of February 2023, the United States' court is currently under pressure of this issue with the case Gonzalez vs. Google, where Gonzalez alleged that Google is liable for the death of their daughter (who died in a terrorist attack in Paris, France, in 2015) because it used computer algorithms to recommend videos published by ISIS or related to ISIS to its users (Barnes et al., 2023). Bertram Lee Jr. of Future Privacy Forum states that "transparency is the key to learning," to know exactly what is being exposed to citizens online and why is a crucial strategy to disseminate harmful information online.

The consequences of algorithmic amplification is the fact that it is based on one's interactions with the platform, their personal interests, and based on what is trending. Therefore, streamed videos of terrorist attacks in real-time will be pushed onto the feed regardless of its harm because it's currently trending. The algorithm is unable to detect the harm content this quickly. Lastly, your personal interaction with content will create filter bubbles, grouping you with like-minded individuals; in this case, extremists.

Evidence on Potential Solutions

Efforts of Social Organizations

Literary experts, government officials, and researchers who work for big tech companies have been trying to figure out whether the algorithms used by social media platforms are making extremist views more popular and leading to increased radicalization. However, social organizations and other third-party groups are often the ones who can bring together all the different viewpoints without any bias. Algorithmic amplification is when some types of content become more popular than others because of the way the algorithm works. This algorithm is powered by the data generated by our online activity such as clicks, likes, comments, and shares. This data is then used to create personalized feeds for users. This section will discuss the current research and findings from unfiltered social organizations.

YouTube is a commonly used platform in research because it has an API that is easy for researchers to use (Thorley, 2022; O'Callaghan, 2015; Ledwich, 2019; Ribeiro, 2019). One research study looked at the impact of YouTube's algorithm on radicalization and found that only one of the four claims was partially supported. The claim that radical bubbles were formed due to the algorithm was found to be partially true because the recommendations provided by YouTube's algorithm stayed within the same categories as the original content viewed by the user. The study showed that even if users were watching extreme content, their recommendations would still include a mix of extreme and more mainstream content. This suggests that YouTube may actually steer users away from extremist content

instead of leading them towards it (Ledwich, 2019). However, anonymity played a role in repopulating mainstream content, which could be seen as a limitation (Whittaker, 2022; Blasiak et al., 2021).

In an interview with Global Internet Forum to Counter Terrorism's (GIFCT) Head of Technology, Tom Thorley expressed that the most tangible results to work from is YouTube, given its accessibility to researchers. The conclusions from YouTube provide tangible and common insights that can be applicable for other social media platforms to build from. For example, when we are able to derive a conclusion like "YouTube's definition of borderline content for TVE is the cause of individual radicalization," we can compare this definition to those of other social media platforms to see how they differ. Data collection and transparency from companies outside of YouTube is rare due to their unwillingness to share such information to the public.

On the other hand, YouTube has also been proven to algorithmically amplify extremist content online (Whittaker, 2021). In a report published in 2022, GIFCT found that most research on this topic uses "black box" testing, where researchers input data and receive outputs without understanding how the underlying algorithms make decisions (Whittaker, 2022). Researchers conducted experiments on YouTube, Reddit, and Gab by creating accounts and following the same channels or subreddits (10 far-right and 10 neutral) (Whittaker, 2021). They left the accounts inactive for a week to create a baseline, and then subjected each account to a different treatment: one interacted with far-right content, one interacted with neutral content, and one remained inactive. The use of automated agents creates behavioral data that the algorithm uses to personalize recommendations. From this practice it was found that the account that predominantly interacted with far- right materials was twice as likely to be shown extreme content, and 1.39 times more likely to be recommended borderline content.

Regarding bots, some social organizations took it upon themselves to use social media bots to promote counter-radicalization (Marcellino et al., 2020). These bots would have online personas that present them as online users of the platform, so that actual online users would assume that they were interacting with another person. The bots would detect when a user would express forms of extremism and hate speech and counter a reply with an attack. The case studies show that bots that interact with

humans one-on-one, as well as vast networks of bots that target whole communities, can empower humans in scalable ways but can also be outmaneuvered by dedicated and intelligent human opponents. This means that most outcomes found that users would still resort to extreme notions of thought, and some users could detect and out-intelligence the bot, which did not make the practice as effective.

From analyzing the successful studies conducted by social organizations, it has been brought to attention that there is no clear definition across all platforms for the terms that are used repeatedly. For example, there is a distinction between 'filter bubbles' and 'echo chambers.' 'Filter bubble', which is driven by platforms without the user's deliberate choice, input, knowledge, or consent, is pre-selected; and 'echo chambers,' in which the user chooses to encounter like-minded views and opinions, are self-selected (Wolfowicz et al., 2021). When there are different known facts and definitions placed for each study, it can be assumed that each of these studies hold different thresholds to determine extremism on social media platforms as well.

The previous studies begin their experiments when a user starts following and interacting with channels that promote TVE content. However, it is a global phenomenon that most people use Google as their search engine, which makes searches the first step in finding TVE content. Ahmed et al. created a keyword categorization in order to construct a list of keywords that potentially leads users to extremist content, whether intentionally or otherwise, if entered into a search engine. While their primary source of search engine was Google compared to other social media platforms, they found that searches are the gateway to TVE content and that many high-risk keywords go unchallenged by counter-narratives (Ahmed & George, 2017). Using the same list of keywords in every language makes it easier to compare results across platforms and languages. It is important to use the same keywords so that we can understand how different platforms and languages handle TVE content and how easy it is to find that content using certain keywords. This approach allows for a broader coverage of content from each platform, including content that is not explicitly posted by TVE-related channels.

Efforts of the Government

In content regulation, it's been found most convenient to take down content when it meets the threshold of extremism. It is important to recognize that this has not always been the case, in fact, academics call out 2016 as a turning point in which platforms began to take a more proactive approach toward removing content and discourse (Rowa, 2022; Ardern, 2022). Scholars have raised concerns about the removal of content and its impact on free speech and transparency. The UK Online Harms White Paper initially recognized the issue of content amplification, but subsequent consultations have downplayed its importance in favor of content removal. The German NetzDG, on the other hand, only addresses the removal of illegal hate speech content and ignores the amplification of extremist content. Both laws aim to regulate harmful content on social media platforms and impose fines if they fail to remove it through notice and takedown mechanisms (Whittaker et al., 2021).

In the growing concern that private social media companies have separate methods of conduct that are held accountable in the form of fees of noncompliance from the government, New Zealand is one of the only countries to address this disconnect with the Christchurch Call Initiative. The disconnect relates back to the limitations from incongruent thresholds and definitions. It is argued that policy is yet to fully understand the difficulties with "grey area" and "borderline" content as it relates to content amplification (Heldt, 2020). The initiative directly won't tell the public all about the outcomes that algorithms are driving online, but it will help with better access to data so researchers can answer these questions (Ardern, 2022). Independent researchers are no longer subjected to only derive from YouTube's research-friendly API.

Efforts of Big Tech

In similar, yet different branding, social media companies have adapted the practice of "preach, but no reach." In 2015, Reddit was the first public social media domain to announce an alteration to recommendations where it was a policy of "quarantining" subreddits. Quarantined subreddits do not appear in non-subscription-based feeds (such as Reddit's "Popular" feed) and are not included in

search or recommendations. This approach is taken to "prevent its content from being accidentally viewed by those who do not knowingly wish to do so or viewed without appropriate context. Whittaker et al.'s study found that Reddit's "Best" timeline did not recommend extreme content (Whittaker, 2022), which could be a result of having removed problematic content due to "quarantining."

Similar to Reddit's 2015 practice, YouTube adopted a tactic of 'reducing' in 2017 as a counter-terrorism policy, where it would take a tougher stance on videos that do not clearly violate policies but may be extreme by removing content from being recommended. According to YouTube, this step reduced views of such videos by an average of 80% (YouTube, 2019). When the policy was expanded to misinformation and conspiracy theories, there was a drop of 70% in views of this content (YouTube, 2020). By a chain effect, Facebook joined the bandwagon by adding three policies: (1) Removing violative content, (2) Reducing misleading content via ranking and (3) Informing users with additional context. This made it apparent that when problematic content that does not violate policies, it can still be harmful to users, and when identified was down ranked in the platform's News Feed. However, Facebook only went beta with this policy for one year when it put countries at risk of conflict (Whittaker, 2022).

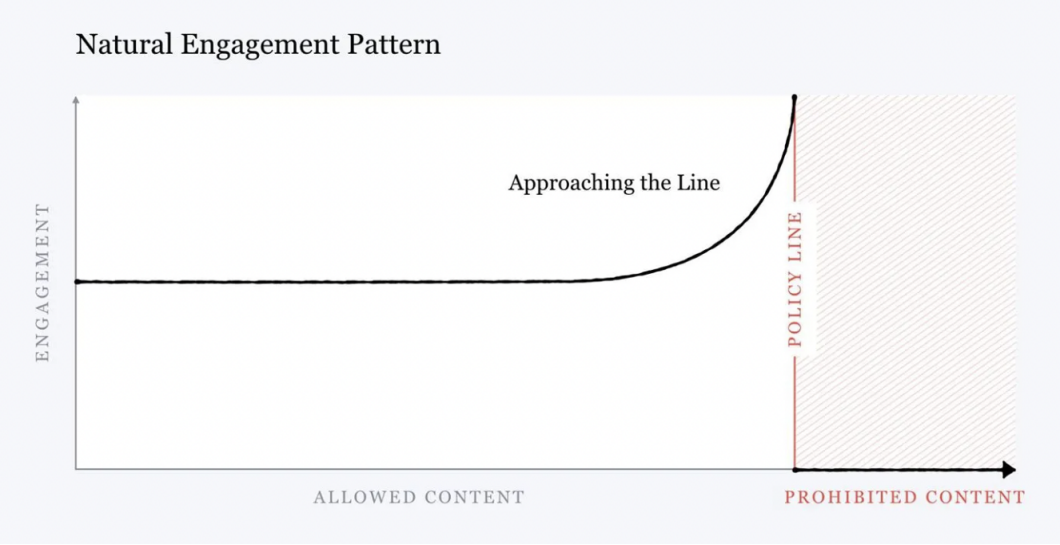


Figure 4: Social Media Companies' Natural Engagement Pattern with content

Source: Constine, 2018. TechCrunch.

Policy Alternatives

Three policy alternatives were chosen based on academic literature, jurisprudence, stakeholder interviews, and cases of terrorism and extremism on social media platforms to tackle the issue of extremist content and promotion by platforms. The policy alternatives to consider for this issue are the maintenance of the status quo, an establishment of baseline definitions for TVE and borderline content, algorithm modification, and the rein on European Union enforcement outcomes.

Maintenance of the Status Quo

The status quo of this issue would be deemed as self-regulation. Self-regulation is an approach where social media companies take responsibility for regulating their own content and operations, without relying on government intervention. However, introduced in November 2022, and soon to be applied in January 2024, the EU has announced the Digital Services Act (DSA) to better protect European citizens from hate speech, disinformation, and other harmful online content. Given the DSA has no historical precedent, the status quo can be considered a fresh alternative. In the context of reducing terrorism and violent extremism content online, the EU would continue to allow social media companies to self-regulate, however, with caution, as now they would have a baseline of restrictions to abide by. Given that the status quo is a new standard, this project treats the status quo as an alternative to build from rather than to replace completely.

The DSA would involve the EU regulating and monitoring social media companies to reduce terrorism and extremism content. It results in stronger enforcement of current policies and the government being responsible for ensuring compliance. The maximum fine for breach of the DSA will be 6% of global annual turnover in the preceding financial year (Kohne et al., 2022). This means that if a company fails to comply with the DSA and is found guilty of a breach, it could be fined up to 6% of its total revenue for the previous year. The European Board for Digital Services: an independent advisory group, would be responsible for enforcing the regulations and monitor compliance by social media companies.

Establishing baseline definitions for TVE and Borderline Content

Algorithms promote legal, yet borderline content that can be harmful and lead to radicalisation (Whittaker, 2020). Borderline content is content that "does not violate the community standards" but is identifiably problematic in an area covered by these "community standards." Across the five platforms my client is covering (Facebook, Instagram, Twitter, YouTube, and TikTok), it's been found that there are different cases of borderline content in TVEC as a result of heterogeneous definitions. Some platforms have taken steps to remove potentially extreme content from their recommendations. For instance, YouTube has announced that content that is deemed borderline will be available behind a warning and not eligible for monetization, recommendation, comments, or endorsements (Whittaker, 2021). Similarly, Facebook has adopted this approach by removing content that violates their community standards or is associated with offline movements tied to violence (Facebook, n.d.). Reddit has a policy of "quarantining" subreddits that are grossly offensive, such as r/The_Donald, which has been accused of hosting problematic extremist content (Reddit, 2021). However, some critics have argued that these measures are only applied after negative media attention.

This alternative proposes that social media platforms establish homogenous, baseline definitions for TVE and borderline content so that they may work in congruence to reduce the amplification of harmful content, rather than independently. The following tables were created as suggested policy of action to enforce baseline borderline content based upon research primarily from industry trends, GIFCT, and the UN. Given that the status quo is a relatively new standard, this policy alternative would be an addition to the status quo if implemented, rather than replacing the status quo.

Taxonomy of Borderline TVE Content

	Benign Borderline	Moderate Borderline	Terrorism and Violent Extremism
Displaying of TVE-related keyword or symbol	✓	✓	✓
	OR	AND	
Displaying of TVE-related video	✓	✓	✓
Displaying of TVE call to action			✓

Table 1. This chart displays what content is accepted as Benign Borderline, Moderate Borderline, and Terrorism and Violent Extremism. It should be noted that this is a 1-2-3 outcome score, where Benign Borderline must only have an outcome of one of three descriptions to be Benign Borderline, Moderate Borderline must have two of the three descriptions, and TVE must have all three descriptions.

Benign Borderline	Moderate Borderline	Terrorism and Violent Extremism
Praise • Displaying or speaking about a designated entity • Displaying or speaking about actions of designated entity • No sense of alignment with designated entity **Support** • Displaying or speaking groups of designated entity • Displaying or speaking of where designated entity receives donations from **Representation** • Displaying or speaking about member of designated entity	**Praise** • Displaying and speaking about a designated entity • Displaying and speaking about actions of designated entity • No sense of alignment with designated entity **Support** • Displaying and speaking groups of designated entity • Displaying and speaking of where designated entity receives donations from **Representation** • Displaying and speaking about member of designated entity	**Praise** • Speaking positively about a designated entity • Give designated entity a sense of achievement • Legitimize their cause by making claims that are hateful, violent or criminal conduct • Alignment with designated entity **Support** • Call for donations or financial support • Material aid or donations • Call for action • Recruitment • Channeling information or resources **Representation** • Claims that make a user member of designated entity • Pages or channels that represent designated entity

Table 2. This table expands on Table 1 by listing examples of praise, support, and representation. The largest difference between Benign Borderline and Moderate Borderline is that Benign Borderline can display or speak, while Moderate Borderline must do both. "Entity" in these charts would be TVE leaders and organizations.

Algorithm Modification

Algorithm modification refers to the process of making changes to the algorithms and recommender systems used by social media companies (Narayanan, 2023), with the objective of reducing the promotion and amplification of TVE content. The process of algorithm modification will be led by social media companies themselves, in response to regulations and frameworks established by governing bodies like the European Commission. Given that the status quo is a relatively new standard, this policy alternative would realistically be an addition to the status quo if implemented, rather than replacing the status quo.

Currently, the only visibility that the EC will have on social media algorithms is through the DSA's required transparency reports, which will be visible once the DSA is enacted in 2024 (Griffin, 2022). Prior to this, only external research organizations conducting projects could determine few insights to algorithms that amplify TVE ideologies (Whittaker, 2020). The NIST AI Risk Management Framework (AI RMF) has been created in partnership with both public and private sectors to enhance the management of risks related to artificial intelligence (AI) for individuals, organizations, and society. This framework serves to be voluntary and aims to enhance the capacity to integrate trustworthiness factors into the design, development, utilization, and assessment of AI services, systems, and products (NIST, 2023). Influence from transparency reports obligated by the DSA, as well as inspiration from an AI Risk Framework like NIST's can enable the EC to create personalized algorithmic modifications for each social media platform.

The Digital Services Coordinators under the DSA will cooperate within an independent advisory group, called the European Board for Digital Services, which can support monitoring, analyses, reports and recommendations, as well as coordinating the new tool of joint investigations by Digital Services Coordinators (Council of the European Union, 2022).

Industry leaders suggest that an open-ended framework for AI governance is more effective than a step-by-step guide due to the rapidly changing landscape. Patrick McLoughlin, Chief Data Officer for the State of Maryland, proposes a governance framework that allows agencies to tailor their mission to it. While Tom Thorley, from Global Internet Forum for Counter Terrorism (GIFCT), expresses that

there are inherent differences of social media platforms, that personalized algorithm modification would account for equal outcome for the reduction of TVE content online. The risk rating assessment can derive a baseline for technical and non-technical risks associated with recommender systems. Specific mitigation strategies depend on the recommender systems of YouTube, Instagram, Facebook, Twitter, and TikTok which will be shown in the transparency reports. Understanding the interconnections between technical and non-technical factors is the first step in improving the system, noting that companies may use different factors per platform. If selected as the policy alternative, the implementation will provide examples of different ways to modify an algorithm based on risk factors. This modification should use these strategies to identify users who actively search for violent or extremist material. In the long term, algorithms should be trained to detect the behavior of users seeking out extreme videos to aid in the advancement of policy research.

Criteria

The selection of criteria and their respective weights was based on the client's objectives and preferences. A rating system ranging from zero (indicating an unfavorable outcome) to three (indicating a highly favorable outcome) was used to evaluate each criterion. Detailed information regarding the point scale and descriptors for each criterion can be found in Tables 1-5 located in the Appendix.

Effectiveness (30%)

EU Policies that target filter bubbles only refer to the transparency of it to users (Johanyak, 2022). Filter bubbles occur when users are only shown content that aligns with their existing beliefs; thus, oftentimes can lead to radicalization dependent on a user's personalized feed. The issue with this is that average users usually do not understand what to do with this information when it's given to them (Schiffer, 2019). A recent study conducted by Princeton University found that a network that tries to create filter bubbles caused an increase in polarization of 4%. In contrast, a network that was left untouched and unregulated increased polarization by a significant 4000% (Chitra & Musco, 2020). This suggests that unregulated networks can have an alarming impact on polarization, which could be

a cause for concern. There is a notable polarization and fragmentation of the political sphere in many countries, including the rise in populist politics and mass protests against immigration in Europe, the polarizing election campaign of Donald Trump, or the British people's vote for Brexit (Geschke, 2018). Therefore, it is not enough for users to know why such content is being recommended to them, but to remove the filtering of extremist content completely. Policies will be rated on their potential to address the existence of extremist filter bubbles to the action of removing them.

Political Feasibility (30%)

Policies are evaluated on political cooperation and support from the European Commission, and public awareness and support on the policy.

Corporate Feasibility (20%)

Policies are evaluated on the capacity and resources to implement, the degree of collaboration and cooperation, and the timeframe for implementation and scalability.

Transparency (20%)

Policies will be evaluated by their ability to remain transparent with users and government agencies. Transparency is a key framework of artificial intelligence that allows humans to see whether the models have been thoroughly tested and make sense, and that they can understand why particular decisions are made (Larsson & Heintz, 2020; Zuiderveen Borgesius et al. 2016). These will be measured on a scale on how comprehensible the information is provided to the user, as well as how much accountability is given to the government.

Criteria for Evaluation

Maintenance of the Status Quo

Effectiveness: Moderately Favorable (2)

The DSA addresses the problem of filter bubbles by mandating transparent disclosure of the functioning of all ranking systems used on platforms such as recommendations and news feed. In case there are multiple recommendation systems in place, users should have the freedom to select the one they prefer (Hildebrandt, 2022). With better transparency with recommendations, users can know what the algorithm groups their interests as. Exposure to this can either prompt users to critically think about their actions, or further reinforce their ideals with confirmation bias. Exposure without the action of reducing filter bubbles can still leave users vulnerable to extremism (Interview with Dr. Steven L. Johnson, 2023). Thus, the DSA builds upon transparency of the potential existence of filter bubbles for users to critically evaluate on their own, but the DSA does not implement a mechanism in which algorithms must reduce personalization to limit the frequency of filter bubbles.

Political Feasibility: Highly Favorable (3)

Adopting the status quo requires no additional cooperation and support from the EU, or public awareness and support, so it is highly favorable and feasible. During the negotiation process, there were early indications suggesting that the final political agreement incorporates certain enhanced provisions that will be crucial in safeguarding freedom of expression (Allen, 2022). Although it will not be enforced until January 2024, the EU is presently in agreement to proceed with the act. International social organizations focused on consumer friendly technology describe the DSA as moving society towards an online world that better respects human rights by effectively putting the brakes on Big Tech's unchecked power (Amnesty International, 2022).

Corporate Feasibility: Favorable (2.66)

The EU is providing a set of standards, conditions, and safeguards. These requirements are not specific to particular pieces of illegal content, but are more systemic. Typically social media companies already

have their own self-regulating frameworks, these regulations would test if their framework comply. While it may have additional costs towards implementation, the frameworks don't have to change a lot in order to comply (The MarkUp, 2022). If companies do not adhere to the recommendations of the status quo, the maximum fine will be 6% of global annual turnover in the preceding financial year (AkinGump, 2022). Companies must determine the tradeoffs internally before deciding to adhere. There are tradeoffs for them to consider on whether their technology will bring it more revenue to continue independently or with the EU. For this reason, companies are on the fence of the Digital Services Act because they believe it will demote a competitive market for their platform. Social media companies face an accelerated compliance timetable for the DSA as it provides them with just four months for this once a designation is made by the Commission (Lomas, 2022).

Transparency: Moderately Favorable (1.5)

There is no historical trend to accurately depict the transparency model of the status quo, given the recent introduction of the DSA Act. Article 29 of the DSA grants a considerable level of discretion to social media platforms in determining the parameters that can be altered or impacted by users, effectively recognizing these platforms as the principal regulators of recommender systems (Whittaker, 2021). Since it is the first iteration of the Act, the assumption can be made that there is some clarity and comprehensibility about the recommendations in a user-friendly format, but users may still not understand what this information means for them (Schiffer, 2019). With more iterations, this can become better over time. Currently, the DSA makes social media companies be transparent to what they are exposing to users; however, they are ultimately in control and accountable for what information is exposed to the public rather than the EU.

Establishing baseline definitions for TVE and Borderline Content

---

Effectiveness: Highly Favorable (3)

Definitions for what is considered borderline content will provide better measures in preventing the creation of TVE- related filter bubbles, assuming that the definition for borderline content will be used to remove potential TVE content online.

Oftentimes, social media companies will consider terrorism and violent extremism as its own definition, and straggle cases become "gray areas cases." These gray areas cases continue to live online because their lack of definition does not allow enough justification to be taken down from the platform. Gray area cases have the ability to form filter bubbles as they relate to terrorism and violent extremism content, thus leading users to harmful pages and communities that will radicalize these ideas further (Whittaker, 2020). TrustLab's experiment in testing the different algorithms of social media platforms for terrorism and violent extremism content has shown that recommendations for TVE content increase the more a user interacts with the content. By the implementation of definitions into an already existing framework, social companies can be instructed to make clear mitigation strategies to take down the potentially radicalizing content online (Constine, 2018). This directly influences the creation of TVE filter bubbles as limiting the visibility to potential TVE content can decrease the interaction with users overall.

Political Feasibility: Favorable (2)

The DSA has already been introduced and is scheduled for implementation in 2024, subject to final stakeholder approval. Given that the European Commission prioritizes user safety against terrorism and has committed efforts towards the DSA, it is highly probable that the establishment of baseline definitions will receive further support. The EU has expressed interest for clear definitions for extremist networks in the EU Parliament Debate earlier this year (CNN, 2023). There is no EU-wide legislation against online hate speech aside from the DSA; however, national governments, led by Germany, have taken a tougher approach against companies hosting illegal content (Human Rights Watch, 2018; Khan, 2019).

As for public awareness and support, Tech Against Terrorism has criticized the approach, arguing that discussions about removing legal content from recommendations are misplaced and show a lack of understanding regarding how terrorists use the internet. They have expressed concerns about the potential negative impact on freedom of speech, the rule of law, and extra-legal norm-setting. Tech Against Terrorism asserts that norm-setting should be driven by consensus-driven mechanisms that are

accountable to democratic institutions (Tech Against Terrorism, 2021). Public awareness and support for borderline definitions is mostly from industry experts in the nonprofit space, rather than the general public because the topic of specific definitions does not reach mainstream media to gain an opinion. Exposure for definitions aims to increase better education for extremist content online and encourage critical thinking users.

Corporate Feasibility: Moderately Favorable (1.42)

As mentioned, popular social media companies such as Facebook, YouTube, and Reddit have implemented measures to take down or demote borderline content on their platforms (Walker, 2017; Facebook, n.d., Reddit, 2021). Majority of the platforms already have sufficient capacity and resources to implement this policy. By the assumption that this policy recommendation would be implemented under the DSA, collaboration is necessary due to regulatory obligation. However, cooperation with definitions can also lead to barriers such as a reduction in competition, which is not easily favored by social media companies (Cusumano et al., 2021). The difficulty with definitions is that the social media companies will hold different thresholds and interpretations of the definitions, thus in the long-term not becoming truly baseline and will continue to cycle of self-regulation. In order to implement definitions for all social media platforms to comply with would require more than a year to follow through, which may lead to limitations and delays dependent on corporate cooperation. This is the assumption given that this is typically the turnaround time for government regulations on social media companies (Lee et al., 2019).

Transparency: Highly Favorable (3)

Building from the assumption that the creation of homogenous borderline definitions will be incorporated into the DSA Act in its later iterations. There is a high probability that there will be better, clear, and comprehensible information of recommendations to a user. This is due to the idea that the reduction of borderline content will allow for potentially harmful content to be less visible to the public; thus, making their moderation processes more transparent to users. The hope for clear definitions is that more gray area cases will be caught to demote it off the platform.

The EU Counter-Terrorism Coordinator has also emphasized the importance of having consistent metrics on borderline definitions for terrorism and violent extremism content through the DSA. This will allow social media companies to provide more detailed information in their transparency reports, including data on their practices for removing illegal and borderline content and whether the content was promoted by their algorithms (Council of the European Union, 2020). The status quo only allows for the transparency reports to be given to the EU. The transparency reports will further allow for the government to have clear accountability and oversight mechanisms to uphold this policy of baseline definitions.

Algorithm Modification

Effectiveness: Highly Favorable (3)

Algorithm modification would inherently address the existence of TVE- related filter bubbles, because it would allow the EU to build from the transparency reports to modify the algorithm to reduce its existence. Algorithmic modification will allow for the identification of TVE content so that the government has visibility into the process, while also acting upon reducing the occurrence of it (Pimental, 2021).

Political Feasibility (3)

Algorithm modification would build upon the existing status quo of the DSA and AI Act in the EU; therefore, it can be assumed that there is strong political will and support for the policy recommendation. The EU AI Act is applicable to all artificial intelligence innovations, while the DSA works specifically within social media companies. The EU already faces consensus that a risk-based approach to assess AI is necessary (European Union, 2022). Civil liberties have assessed that the NIST AI Risk Management Framework has a more flexible approach in assessing social media platforms which will allow the platform to remain user-enticing given their algorithm will still remain within its original intent (Brookings, 2023).

Corporate Feasibility: Favorable (2.33)

As mentioned, popular social media companies frequently change their algorithms in order to keep up with the demands of society. Similar to the understanding of establishing baseline definitions, social media platforms already have sufficient capacity and resources to implement this policy since they already have been practicing it internally. By the assumption that this policy recommendation would be implemented under the DSA, collaboration is necessary due to regulatory obligation. Opposite to the establishing baseline definitions, there will be a high level of cooperation from social media platforms as algorithmic modification does not eliminate economic competition between them. Modifications allow for personalized approach that leads to the outcome of reduce TVE content online, but does not morph all the algorithms to be exactly the same (Kerry, 2023).

Under similar circumstances of establishing a baseline definition, the timeline would require more than a year to enact the policy and modify the algorithm (Lee et al., 2019); however, there is a greater lack of timeframe given that it is a personalized approach towards modification.

Transparency: Favorable (2.5)

It can be anticipated that if the DSA Act incorporates personalized modification of algorithms in its future iterations, users may receive clearer and more understandable information regarding recommendations. However, it should be noted that this is a forward-looking perspective reliant on the status quo, and modifying algorithms itself does not necessarily guarantee an increase in transparency for the user (NIST, 2023). Algorithmic modification mainly serves as transparency to the government, given the highly classified nature of social media algorithms. The EU will use their transparency reports to develop modifications to each respective algorithm. This will demand strong accountability and oversight strictly from the government to modify noticeable harms (Kerry, 2023).

Recommendation

Based on the projected outcomes of policy alternatives, I recommend that Trust Lab advocates for the European Commision to incorporate algorithm modification to reduce terrorism and violent extremism content online, which will reduce radicalization to real world harm. This policy option received highly favorable on effectiveness, highly favorable on political feasibility, favorable on corporate feasibility, and favorable on transparency.

Outcomes Matrix

	Status Quo	**Establishing baseline definitions for TVE and Borderline Content**	**Algorithm Modification**
Effectiveness	Moderately Favorable (2)	Highly Favorable (3)	Highly Favorable (3)
Political Feasibility	Highly Favorable (3)	Favorable (2)	Highly Favorable (3)
Corporate Feasibility	Favorable (2.66)	Moderately Favorable (1.42)	Favorable (2.33)
Transparency	Moderately Favorable (1.5)	Highly Favorable (3)	Favorable (2.5)
Overall Score	Favorable (2.332)	Favorable (2.384)	Favorable (2.766)
Equal Weights	Favorable (2.29)	Favorable (2.355)	**Favorable (2.7)**

Implementation

Technology is outpacing governance is the common phrase one hears when debating for better regulations for social media companies and technology overall. Though innovation for social media companies and recommender systems cannot all be predicted in advance, if policymakers make feedback anticipation a critical element in the design process, it can incorporate visibility from all perspectives to compromise a plan. Algorithm modification refers to the modification of the algorithms used in recommender systems to promote mitigation strategies such as fairness, transparency, and accountability in the content they recommend. The policy recommendation of algorithm modification for recommender systems aims to tailor the mission of social media companies to the regulatory framework, taking into account different sizing, scalability, and resource availability of each one. Artificial Intelligence within technology is a growing industry that has limited regulation; however, the National Institute of Standards and Technology in the U.S. has very recently developed an AI Framework Playbook that will assess technology based upon the risks they pose (NIST, 2023). This proves the growing concern for regulation over algorithms, as well as this framework serving as a first step towards safer innovation.

It must be understood that all the components used in a recommender system have the potential to significantly affect user experience and cause harm. Nonetheless, it's important to recognize that such factors are not absolute and can have complex interdependencies. In an interview with Dr. Steven L. Johnson, co-author of "Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumptions," has identified that findability and discoverability are key points for algorithms to detect to reduce the escalation of radicalization in motivated users. Findability refers to how easy it is to find specific content on a social media platform, both on and off the platform. It is closely related to discoverability, which is the process of discovering content that is recommended to the user. Findability depends on the user's own searching and interest, while discoverability is based on content that is recommended by the platform. It's important to understand that high findability of certain content can lead to high discoverability, depending on the

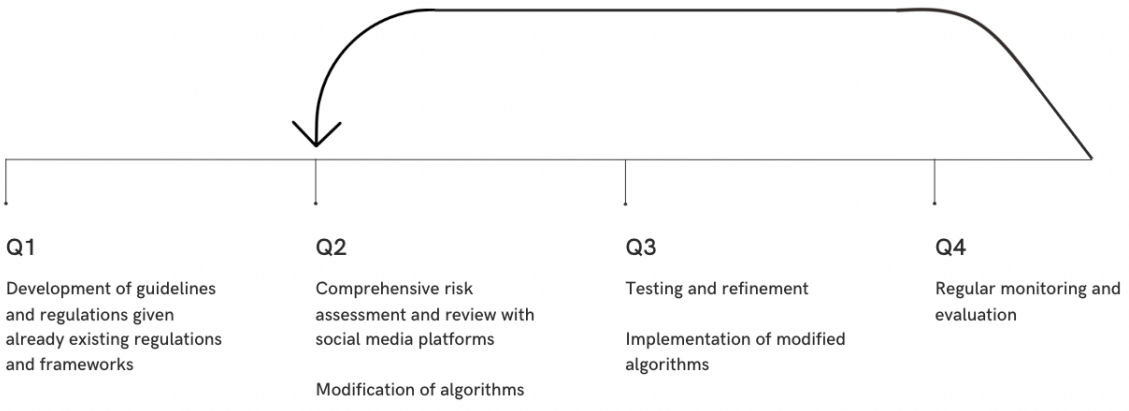
platform's algorithm. For example, if an individual frequently searches for and interacts with violent extremist content, they may be recommended more of that content.

If we use the example of findability to determine algorithmic modification for social media platforms, a potential mitigation strategy, specifically the ones that increase in findability for TVE-related content, is to modify their algorithms to intervene with a motivated user's search to stop showing them new content as it relates to TVE the more time they spend searching. This intervention should prevent the platform from showing the user new TVE-related content as they continue to search. By repeatedly showing the same, and not explicitly harmful information to the user from their initial search, their escalation in radicalization will decrease over time, as they do not receive new information. This would require the project finding from Trust Lab to determine which social media companies have high findability in their algorithms. Once findability is understood, then the different factors of a recommender system must be detected to reduce overall findability for TVE-related content.

The stakeholders involved in moving the recommendation forward include social media companies, regulatory agencies, and user groups. Social media companies play a crucial role in implementing the new algorithms, while regulatory agencies provide oversight and enforce compliance (Hananel et al., 2021). Given that social media companies are currently not as transparent about the explainability of their algorithms, Trust Lab is projecting how much TVE-related content lives on each platform and to what extent. Better visibility into this will allow for algorithm modification that is tailor to each company. User groups, on the other hand, are the primary beneficiaries of algorithm modification for recommender systems, as they can benefit from fairer and more transparent content recommendations.

The following steps are necessary to move the recommendation within a year (4 quarters):

1. The current European Board for Digital Services, an independent advisory group, would be responsible to develop guidelines and regulations for algorithm modification. (AkinGump, 2022). The task force consists of experts from different fields, including computer science, law, policy, ethics, and social sciences. The team will collaborate to develop a set of guidelines and regulations for algorithm modification for recommender systems.
2. Conduct a comprehensive risk assessment of current algorithms and review them against the guidelines and regulations. This will take cooperation from social media companies to be more transparent and accountable on their recommender systems. Trust Lab is currently in the process of developing the distinction between each social media algorithm through private research.
3. Modify algorithms to comply with the mitigation strategy relevant to the respective social media company for the overall goal of reduce TVE- related content online.
4. Test and refine the modified algorithms to ensure their effectiveness and compliance.
5. Implement the modified algorithms in the recommender systems of social media companies.
6. Monitor and evaluate the modified algorithms regularly to ensure their compliance with the guidelines and regulations.



Conclusion

The radicalization from terrorism and violent extremism content online has been present across the world, and in Europe. Horrific events like the Lockerbie bombings and Christchurch shooting are grave reminders that individuals are capable of fatal harm. In order to reduce radicalization in Europe, algorithmic amplification must be targeted. Through extensive research and European Union jurisprudence, as well as stakeholder analysis, it has come to the conclusion that algorithmic modification is feasible to reduce terrorism and violent extremism content online, and radicalization in the long-term.

Appendix

Table 1: Effectiveness Projection Table

	Low (+1)	Medium (+2)	High (+3)
Filter Bubbles	Does Not Address	Moderately addresses by means of giving transparency to the user and/or government	Significantly addresses by giving government ability to suggest modifications

Table 2: Political Feasibility Projection Table

	Low (+.5)	Medium (+1)	High (+1.5)
Political Cooperation and Support (EU)	Lack of political will or support for the modification	Some political will and support, but with significant barriers or challenges	Strong political will and support for the policy
Public Awareness and Support	Low public awareness or support for the modification	Some public awareness or support, but with significant opposition or resistance	Strong public awareness and support for the policy

Table 3: Corporate Feasibility Projection Table

	Low (+.33)	Medium (+.66)	High (+1)
Capacity and resources	Limited capacity or resources to implement the modification or technology	Adequate capacity and resources, but with some limitations or challenges	Sufficient capacity and resources to implement the modification or technology
Collaboration and cooperation	Little or no collaboration or cooperation with external stakeholders	Some collaboration and cooperation, but with some barriers or challenges	High level of collaboration and cooperation with external stakeholders
Timeframe for implementation and scalability	Limited timeframe for implementation or scalability	Moderate timeframe for implementation or scalability, with some potential for delays or limitations	Sufficient timeframe for implementation and scalability

Table 4: Transparency Projection Table

	Low (+.5)	Medium (+1)	High (+1.5)
Clarity and comprehensibility	Lack of clarity or comprehensibility	Some clarity and comprehensibility in the information provided, but with some gaps or ambiguities	Clear and comprehensive information provided in a user-friendly format
Accountability and oversight	Lack of accountability or oversight mechanisms from EU in place	Some accountability mechanisms from EU in place, but no accountability through modification	Strong accountability and oversight mechanisms in place to modify existing harms

Table 5: Numeric to Descriptor Indicator

Descriptor	**Highly Unfavorable**	**Unfavorable**	**Moderately Favorable**	**Favorable**	**Highly Favorable**
Number	0	0 < X <= 1	1 < X < 2	2 <= X < 3	3

Bibliography

Ahmed, M., & George, F. L. (2017). A war of keywords: How extremists are exploiting the internet and what to do about it. Center on Religion and Geopolitics.

Allen, A. (2022, April 26). The Digital Services Act: Political agreement reached, long road ahead awaits. Center for Democracy and Technology. https://cdt.org/insights/the-digital-services-act-political-agreement-reached-long-road-ahead-awaits/

Amnesty International. (2022, April 26). European Union: Digital Services Act Agreement a 'watershed moment' for internet regulation. Retrieved from https://www.amnesty.org/en/latest/news/2022/04/european-union-digital-services-act-agreement-a-watershed-moment-for-internet-regulation/

Ardern, J. (2022, September 21). Christchurch Call Initiative on Algorithmic Outcomes. The Beehive. Retrieved October 19, 2022, from https://www.beehive.govt.nz/release/christchurch-call-initiative-algorithmic-outcomes

Barnes, R., Vynck, G., Lima, C., Oremus, W., & Wang, A. (2023, February 21). Supreme Court considers if Google is liable for recommending Isis Videos. Retrieved from https://www.washingtonpost.com/technology/2023/02/21/gonzalez-v-google-section-230-supreme-court/

BBC. (2020, August 24). Christchurch shooting: Gunman Tarrant wanted to kill 'as many as possible'. Retrieved from https://www.bbc.com/news/world-asia-53861456

Bécuwe, N., Goudet, S. & Tsoulos-Malakoudi, D. (2018) Survey report 'European youth and radicalisation leading to violence': analysis and recommendations for policy-making purposes. https://www.precobias.eu/wp-content/uploads/2021/06/PRECOBIAS-toolkit-definitive-edition.pdf)

Blasiak, K. M., Risius, M., & Matook, S. (2021). "Social Bots for Peace": A Dual-Process Perspective to Counter Online Extremist Messaging.

Britannica. (2023, February 13). Pan Am Flight 103. Retrieved from https://www.britannica.com/event/Pan-Am-flight-103

Canlas, K. (2019, March 21). Calls for social media firms to take responsibility over violent content. Retrieved from https://m.mpamag.com/nz/news/general/calls-for-social-media-firms-to-take-responsibility-over-violent-content/305381

Chan, K. (2022, April 23). EU law targets Big Tech over hate speech, disinformation. AP NEWS. Retrieved from https://apnews.com/article/technology-business-police-social-media-reform-52744e1d0f5b93a426f966138f2ccb52

Chitra, U., & Musco, C. (2020, January 1). Analyzing the impact of filter bubbles on social network polarization: Proceedings of the 13th International Conference on Web Search and data mining. ACM Conferences. https://dl.acm.org/doi/10.1145/3336191.3371825

CNN. (2023, January 19). EU debates implications of Far right 'terror networks' news | EU parliament Debate | News18 Live. YouTube. https://www.youtube.com/watch?v=qqX96ntb1UM

Constine, J. (2018, November 15). Facebook will change algorithm to demote "borderline content" that almost violates policies. Retrieved from https://techcrunch.com/2018/11/15/facebook-borderline-content/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEf-lKgQ6TfWcWoz5PwJwf8ce6qFCc65JkwWObD_y9lhf_kfzPK77ichyhfVk0J0HGapFJkAbsSSe_RJqn7goRfpSBS8Eijx11OiVuANkqwPO8FNohgx-OTaiUdN5AE4E4BowvwJsjmxh3UDSDDv2coT_Nx2BPg2Hl97_yWBg6D0

Cusumano, M. A., Gawer, A., & Yoffie, D. B. (2021, January 19). Social media companies should self-regulate. now. Harvard Business Review. https://hbr.org/2021/01/social-media-companies-should-self-regulate-now

Dang, S. (2022, October 28). As Elon Musk takes over Twitter, free speech limits tested. Reuters. Retrieved from https://www.reuters.com/technology/elon-musk-takes-over-twitter-free-speech-limits-tested-2022-10-28/

Doxsee, C., Jones, S., Thompson, J., Halstead, K., & Hwang, G. (2022, May 17). Pushed to extremes: Domestic terrorism amid polarization and protest. Retrieved from https://www.csis.org/analysis/pushed-extremes-domestic-terrorism-amid-polarization-and-protest

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.

European Commission. (2023, April 25). Questions and Answers: Digital Services Act. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348

European Union: Digital Services Act Agreement a 'watershed moment' for internet regulation. Amnesty International. (2022, April 26). https://www.amnesty.org/en/latest/news/2022/04/european-union-digital-services-act-agreement-a-watershed-moment-for-internet-regulation/

EU strengthens control of the acquisition and possession of firearms. (2017, April 25). Retrieved from https://www.consilium.europa.eu/en/press/press-releases/2017/04/25/control-acquisition-possession-weapons/

Facebook Help Centre. (nd). What are recommendations on Facebook? https://www.facebook.com/hel p/1257205004624246

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

Flynn, M. (2019, March 19). No one who watched New Zealand shooter's video live reported it to Facebook, Company says. Retrieved from https://www.washingtonpost.com/nation/2019/03/19/new-zealand-mosque-shooters-facebook-live-stream-was-viewed-thousands-times-before-being-removed/

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. British Journal of Social Psychology, 58(1), 129-149.

Gilbert, D. (2022, October 28). Neo-Nazis, antisemites, and the N-word: Twitter just hours under Elon Musk. VICE. Retrieved from https://www.vice.com/en/article/jgpkqb/elon-musk-twitter-neo-nazis

Griffin, R. (2022, August 02). Does the Digital Services Act have anything to say about the 'tiktokification of Instagram'? Retrieved April 26, 2023, from https://techpolicy.press/does-the-digital-services-act-have-anything-to-say-about-the-tiktokification-of-instagram/

Hananel, S., Gordon, P., Fowler, N., McConville, D., Jarsulic, M., Sutton, T., & Zhavoronkova, M. (2021, November 19). How to regulate tech: A technology policy framework for online services. Center for American Progress. https://www.americanprogress.org/article/how-to-regulate-tech-a-technology-policy-framework-for-online-services/

Heldt, A. (2020). Borderline speech: caught in a free speech limbo?. Internet Policy Review, 15.

Hildebrandt, M. (2022, April 28). The issue of proxies and Choice Architectures. why EU law matters for Recommender Systems. Frontiers in artificial intelligence. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9096719/

Human Rights Watch. (2020, October 28). Germany: Flawed social media law. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

ICO. (2018, August 02). Guide to the UK General Data Protection Regulation (UK GDPR). Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

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

Johanyak, C. (2022). EU to burst the social media bubble. Retrieved from https://hypeandhyper.com/eu-to-burst-the-social-media-bubble/

Kerry, C. F. (2023, February 15). NIST's AI Risk Management Framework plants a flag in the AI debate. Brookings. https://www.brookings.edu/blog/techtank/2023/02/15/nists-ai-risk-management-framework-plants-a-flag-in-the-ai-debate/

Khan, M. (2019, February 4). More 'hate speech' being removed from social media. Subscribe to read | Financial Times. https://www.ft.com/content/868f9d96-27bc-11e9-a5ab-ff8ef2b976c7

Kohne, N. G., Reed, M. A., Garrod, D., Arlington, J., Gleeson, M., & Armytage, A. (2022, July 20). Digital Services Act: Protecting the digital space against the spread of illegal content. Akin. https://www.akingump.com/en/insights/alerts/digital-services-act-protecting-the-digital-space-against-the-spread-of-illegal-content

Larsson, S., & Heintz, F. (2020, May 5). Transparency in artificial intelligence. Internet Policy Review. https://policyreview.info/concepts/transparency-artificial-intelligence

Ledwich, M., & Zaitsev, A. (2019). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. http://arxiv.org/abs/1912.11211.

Lee, N. T., Resnick, P., & Barton, G. (2019, March 9). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Lomas, N. (2022, November 16). EU's Digital Services Act enters into force - but no confirm if Twitter will feel its full force yet. TechCrunch. Retrieved from https://techcrunch.com/2022/11/16/digital-services-act-enters-into-force/

Mackrael, K. (2022, April 25). European lawmakers reach new deal on social media regulations. Retrieved from https://www.wsj.com/articles/european-lawmakers-negotiating-new-social-media-regulations-11650647701

Malone, I. (2022, May 21). Will the EU's Digital Services Act reduce online extremism? Just Security. Retrieved from https://www.justsecurity.org/81534/will-the-eus-digital-service-act-reduce-online-extremism/

Marcellino, W., Magnuson, M., Stickells, A., Boudreaux, B., Helmus, T. C., Geist, E., & Winkelman, Z. (2020). Counter-Radicalization Bot Research Using Social Bots to Fight Violent Extremism. RAND CORP SANTA MONICA CA.

Narayanan, A. (2023, March 09). Understanding social media recommendation algorithms. Retrieved from https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

NIST AI RMF playbook. NIST. (2023, February 1). https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

O'Callaghan, D. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. Social Science Computer Review, 33(4), 459–478. https://doi.org/10.1177/08944393145553 29

Paeman, D., Savova, D., Kennis, M., Kennedy, A., & Flakoll, R. (2022, September 14). Digital Services Regulation in the EU: An evolving landscape. Retrieved from https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2022/09/digital-services-regulation-in-the-eu-an-evolving-landscape.html

Pariser, E. (2011). Beware online "filter bubbles". Retrieved from https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles

Pimentel, H. (2021, April 20). Should the government play a role in reducing algorithmic bias? Brookings. https://www.brookings.edu/events/should-the-government-play-a-role-in-reducing-algorithmic-bias/

Reddit Help. (2021). Quarantined Subreddits. https://reddit.zendesk.com/hc/en-us/articles/36004306 9012-Quarantined-Subreddits

Ribeiro, M. H. (2019). Auditing Radicalization Pathways on YouTube. ACM Symposium on Neural Gaze Detection. http://arxiv.org/abs/1908.08313.

Rowa, J. Y. (2022, July). The Contextuality of Lone Wolf Algorithms: An Examination of (Non)Violent Extremism in the Cyber-Physical Space. GIFCT.org. Retrieved October 19, 2022, from https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-ContextualityIntros-1.1.pdf

Schiffer, Z. (2019, November 12). 'filter bubble' author Eli Pariser on why we need publicly owned social networks. The Verge. https://www.theverge.com/interface/2019/11/12/20959479/eli-pariser-civic-signals-filter-bubble-q-a

Statista. (2023, February 28). Topic: Terrorism in Europe. Retrieved from https://www.statista.com/topics/3788/terrorism-in-europe/

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

Tech Against Terrorism. (2021, February). Content personalisation and the online dissemination of terrorist and violent extremist content. Retrieved by https://www.techagainstterrorism.org/wp-content/uploads/2021/06/210120-TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content-JB-vFINAL-DA.pdf

The Markup. (2022, April 30). Understanding the digital services act – the markup. Retrieved from https://themarkup.org/newsletter/hello-world/understanding-the-digital-services-act

Thorley, T., Llansó, E., Meserole, C. (2022, July). Methodologies to Evaluate Content Sharing Algorithms & Processes. GIFCT.org. Retrieved from https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TA-Evaluate-1.1.pdf

UNODC. (2018). Counter-terrorism module 2 key issues: Radicalization & Violent extremism. Retrieved from https://www.unodc.org/e4j/zh/terrorism/module-2/key-issues/radicalization-violent-extremism.html

Walker, K. (2017, June). Four Steps We're Taking Today to Fight Terrorism Online [Blog post]. Google. https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-onli ne-terror/

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. Internet Policy Review, 10(2), 1-29.

Whittaker, J. (2022, July). Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence. GIFCT.org. Retrieved from https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf

Wolfowicz, M., Weisburd, D., & Hasisi, B. (2021). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. Journal of Experimental Criminology, 1-23.

YouTube. (2019, September 3). The four RS of responsibility, part 1: Removing harmful content. blog.youtube. Retrieved from https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/

YouTube. (2020, October 15). Managing harmful conspiracy theories on YouTube. blog.youtube. Retrieved from https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/

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

Zuiderveen Borgesius, F. J. & Trilling, D. & Möller, J. & Bodó, B. & de Vreese, C. H. & Helberger, N. (2016). Should we worry about filter bubbles?. Internet Policy Review, 5(1). https://doi.org/10.14763/2016.1.401