GIFCT
Global Internet Forum to Counter Terrorism

Content-Sharing Algorithms, Processes, and Positive Interventions Working Group

Part 1: Content-Sharing Algorithms & Processes

July 2021

GIFCT
Global Internet Forum to Counter Terrorism

Executive Summary

Over the past 12 months, representatives from government, tech, and civil society have come together as part of the GIFCT Content-Sharing Algorithms, Processes, and Positive Interventions (CAPPI) Working Group (WG). The group adopted the shared goal of mapping content-sharing algorithms and processes used by industry that could facilitate consumption of content that may increase user interest in or amplify terrorist and violent extremist content and consider positive interventions and risk mitigation points. This report focuses on the first part of that goal – to map "content-sharing algorithms" and processes used by industry (i.e. algorithms that organize information and power the content-sharing features of services). This multi-stakeholder exercise seeks to inform the reader about industry uses of such systems to increase awareness and understanding. It sets out how certain processes, such as search, recommendation, and ad tech algorithms could be exploited by bad actors. Finally, it identifies knowledge gaps in what is known about terrorist and violent extremist content (TVEC) and algorithmic processes, what data and information may help to address those gaps, and recommendations GIFCT may wish to consider.

Working groups are a multi-stakeholder effort to further discussion on the given topic of the nexus between terrorism and technology. This paper represents a diverse array of expertise and analysis coming from tech, government, and civil society participants. It is not a statement of policy, nor is this paper to be considered the official view of the stakeholders who provided inputs.

Introduction

GIFCT Member companies prohibit terrorist and violent extremist content (TVEC) under their terms of service, community guidelines, or content policies. This is one prerequisite to GIFCT Membership and means that platforms remove this content if and when they become aware of it on their services. However, content removal is just one lever platforms may use when it comes to information quality and content moderation.

Following the Christchurch attack in New Zealand on March 15, 2019, several tech companies and governments joined the Christchurch Call to Action. In joining the call, industry members committed to **"review the operation of algorithms that may amplify terrorist and violent extremist content."** In July 2020, the GIFCT established two WGs focused on algorithms and positive interventions to practically take forward the call commitment in a multi-stakeholder forum. These two WGs were then combined to form one single working group. The CAPPI WG is made up of representatives from governments, tech companies, and civil society, including academia, practitioners, human rights experts, researchers, and members of the NGO community.

In setting out the work program for the past 12 months, the group's shared objective was to collaborate across industry, government, and civil society **"to map content-sharing algorithms and processes used by industry that may facilitate consumption of content that may increase user interest in or amplify terrorist and violent extremist content and consider positive interventions and risk mitigation points."**

This report focuses on the first part of that objective: **To map content-sharing algorithms and processes used by industry**. The potential role of content-sharing algorithms in radicalization and violent extremist recruitment continues to be an issue of focus. Yet there is little understanding about what those algorithms are and how they might be misused, contribute to radicalization, or be exploited by violent extremist groups and bad actors. This document is the first step in contributing to that understanding.

In the sections below, the document walks through examples of content-sharing algorithms, how they might be exploited, and what the current literature suggests about their potential impact. While the GIFCT's mission is focused on preventing terrorists and violent extremists from exploiting member platforms, for the purposes of this work we have taken a content-agnostic approach and sought to map the content-sharing algorithms and recommendation processes used across the online ecosystem. This is not with any intention of drawing conclusions or establishing correlations between the uses of content recommendation systems and amplification of any particular type of content. Rather, this mapping exercise seeks to inform the reader about the industry uses of such systems to increase multi-stakeholder awareness and understanding of frequently misunderstood processes. It sets out how these processes, such as search,

recommendation, and ad tech algorithms, could be exploited by bad actors. Many platforms already take mitigating steps to prevent such potential exploitation. One important step is the use of "positive interventions." More information about positive interventions can be found in the second part of this WG report (CAPPI Part 2: Positive Interventions).

Finally, this paper identifies knowledge gaps in what is known about TVEC and algorithmic processes, what data and information would help to address those gaps, and what linkages exist between our understanding of the impact of content-sharing algorithms and the potential for positive interventions.

I. Overview of Content-Sharing Algorithms

Content-sharing algorithms come in a wide variety of forms. While users share content and algorithms organize information, for the purposes of this report, we refer to them as "content-sharing algorithms." This includes web search, newsfeed algorithms, and any algorithm that attempts to organize and curate content in any form (text, audio, or video). However, the impact of content-sharing algorithms is not well understood outside of the tech industry, largely because the underlying algorithms are not either.

This section walks through three broad classes of content-sharing algorithms: search algorithms, recommendation algorithms, and ad tech algorithms.

Search Algorithms

Search algorithms are a particularly effective form of organizing content. Because they rely on explicit user input, search algorithms can make strong inferences about what a given user's intent is, and then use that inference to select the most relevant and/or useful pieces of content in an index and present them to the user. In contrast to recommendation algorithms that rely on user behavior alone to surface content, search algorithms have a stronger signal leverage. The search input's capture of user intent enables search algorithms to be highly efficient at surfacing and sharing relevant content.

Search algorithms are distinguished by several criteria. First and most obviously, search algorithms are distinguished by the dataset they search through (the "index"). Perhaps the most well-known type of search algorithm, web search, typically crawls publicly available websites and returns the sites or URLs most likely to provide the user with relevant information. By contrast, many social networks and digital platforms also offer a form of product search, which typically helps users find content on a service by combing through data collected by the platform and returns the most relevant content. The search function on Pinterest, for example, only returns content shared on its platform.

Second, search algorithms are also distinguished by the type of algorithm used. Today most web search algorithms, as well as the product search algorithms of large platforms, rely at least in part on **machine learning** algorithms. By sifting through and learning from a vast array of data, machine learning algorithms often learn to link concepts together. By contrast, simpler search products will rely on **deterministic** algorithms that match queries exactly. Consider a search for "eye doctor": a search engine based on machine learning will learn on its own to return results for "ophthalmologist" as well as "eye doctor," but a deterministic algorithm won't return results from "ophthalmologist" unless it is explicitly programmed to do so. By linking words and concepts together, search algorithms that use machine learning are generally both more powerful and more useful. Whether and how a search algorithm relies on machine learning thus could have significant implications for the potential spread of content and information related to TVEC (including opportunities for positive interventions).

Finally, as with recommendation algorithms, some web and product search algorithms are now personalized. Rather than rely solely on the input query, some search algorithms now also rely on user data and behavior to surface search results. Most platforms allow users to opt-out of personalization, but the rise of personalized web and product search has important implications for how search algorithms may be exploited by terrorist or violent extremist groups (as well as how they may be leveraged for positive interventions).

Recommendation Algorithms

Many networks and platforms also use recommendation algorithms to curate content they expect their users may be interested in. In contrast to search algorithms, recommendation algorithms typically do not share content in response to explicit user input such as a search query, but instead surface relevant and engaging content automatically.

Personalized recommendation algorithms draw on information about given users and their past behavior on a given platform to recommend items that the platform expects will be most interesting and/or useful to these users. Some of these algorithms are central to a product or service, such as the algorithms that order the newsfeeds of users on Facebook or Twitter. In other cases, however, recommendations are peripheral to an application or website's core product or service, and may be placed on the screen or in an interaction design in a way that the user must deliberately choose to consume (e.g., YouTube's recommended videos toolbar or sidebar). Yet there is one thing all forms of personalized recommendation algorithms have in common: they share content with a user based on the user's known attributes, interests, or behaviors. Although public information provided about specific recommendation algorithms is often limited, one of the parameters of recommendation algorithms is to predict items of content that a user is

likely to find useful based on a wide variety of individual features' data points.[1] Examples abound: recommendation algorithms may draw on the historic behavior of the individual user regarding item features, network interaction, and types of engagement per item type; experimental surveys to understand user preferences; explicit user feedback on the selection of recommended content they receive; explicit stated interests of the user; contextual information (e.g. time of day); and user information, including geography, age, and device type. Given the sheer range of potential inputs, different users engaging with the same algorithm will see a different set of recommended content.

In addition to personalized recommendations, many platforms also rely on algorithms to surface "trending topics" or "most popular" recommendations. Although the algorithms underlying these sections differ from platform to platform, they generally recommend content that has seen a sudden increase in user engagement. For instance, the trending topics section on Twitter may feature a hashtag that has quickly appeared in a high number of posts, while for YouTube it may feature videos that received an unusually high level of user engagement. Almost always, however, trending topics algorithms share content based on recent popularity.

As with all algorithms, the sophistication of trending topics algorithms varies widely. A naive or unsophisticated algorithm will simply count how much engagement every piece of content on the site has gotten over the past hour or day and share the content with the greatest engagement. However, if some content is consistently popular, the trending section will not vary over time. A common way to address this is to leverage machine learning or advanced statistics to establish the normal frequency with which a given phrase or piece of content appears, and then identify the topics or content whose popularity most exceeds what would normally be expected.

Ad tech algorithms

Many popular digital platforms monetize their products via advertising. In contrast to advertising on traditional broadcast and analog media, some (but not all) advertising on digital platforms works by sorting a platform's users by their known interests and demographic information, and then programmatically (and anonymously) matching advertisers to users in real-time as a given web page or piece of content is being loaded. By leveraging data on a user's demographics and behavior, ad tech algorithms can reach specific demographic audiences on their platform without sharing user data with advertisers directly.

Compared with search algorithms and recommendation algorithms, ad tech algorithms are unique insofar as they provide third parties a mechanism for directly reaching

1 Due to their proprietary nature, many platforms disclose only high level details about how they work. For example, see these blog posts by Facebook, YouTube, and Google.

specific audiences. Most ad tech content policies prohibit TVEC. However, where this content isn't moderated, policies can have two important implications. On the one hand, they offer extremist movements a potential way to generate financial resources or reach out to new audiences (as discussed below). On the other hand, they also offer a cost-effective way for researchers and practitioners to carry out positive interventions, as discussed in the CAPPI Positive Interventions Report.

II. Exploitation of Content-Sharing Algorithms

Social networks and digital platforms use a wide range of algorithms to organize content for their users. By surfacing content that users are likely to be interested in, content-sharing algorithms play a valuable role in distributing content and information online. Yet precisely because they can deliver relevant and engaging content to users so effectively, terrorist and violent extremist movements have also sought to take advantage of them. Although content-sharing algorithms are not used to amplify known TVEC material – most networks and platforms remove TVEC as soon as it is discovered – they can spread content that may play a role in terrorist or extremist processes even if the content itself is not illegal or does not violate a platform's terms of service. What to do about the amplification of such content is the subject of significant policy debate, and has led several platforms to restrict the visibility of so-called "borderline" material.

Yet how to respond to the sharing of controversial or borderline content first requires an informed understanding of what its impact may be on violent extremism. Accordingly, this section aims to outline several theoretical mechanisms by which content-sharing algorithms may be exploited (intentionally or unintentionally) by terrorist groups or violent extremist movements.

Search Engines

- **Data Voids:** search engine queries that turn up few to no results, especially when the query is rather obscure, or infrequently searched. Data voids can be exploited to expose people to problematic content. They are often difficult to detect and are mostly harmless until an event causes a lot of people to search for the same term. Golebiewski and Boyd identify five types of data voids: Breaking News, Strategic New Terms, Outdated Terms, Fragmented Concepts, and Problematic Queries. For example, a violent extremist movement could use a data void caused by a breaking news event to spread propaganda by producing and linking to posts and videos titled with a unique search term related to the event that otherwise lacks results. Data voids are often manipulated in tandem with auto-play, auto-fill, and trending topics recommendation algorithms.
- **Search engine optimization (SEO):** the process of improving a website's position in search results so that the webpage features higher on the results page of major

search engines for relevant searches. Generally, this process can be broken into four main steps: keyword research, indexing, on-site optimization, and off-site optimization. SEO methods that violate the major search engines' guidelines are called "black hat" methods. These include the use of keyword stuffing where a target keyword is repeated throughout a webpage, often hidden in HTML elements or stylesheets. Another method is referred to as "cloaking," where different content is provided to the search engines and human visitors. Unlike data voids, the goal with black hat SEO is not to populate results for a rare or infrequently used search term, but instead to manipulate search engine results for popular or commonly used keywords and queries.

- **Autocomplete:** Another possible means of exploitation is through the search engine's autocomplete function. Autocomplete manipulation is an example of black hat SEO. These manipulations take advantage of the way a search engine ranks suggestions for a given trigger. A trigger is the small portion of a search query a user enters before autocomplete provides a set of suggestions identified from common search terms. As this process is based on the popularity of queries observed from search logs, it has been vulnerable to manipulation via spam. While it is possible that bad actors could manipulate autocomplete suggestions to promote violent extremist terminology or materials, it is also possible that these terms could be inadvertently pushed by CVE campaigns.

Recommendation Algorithms

- **Echo chambers:** the more social networks & digital platforms rely on personalized recommendation algorithms to sort and share content with users, the greater potential there is for the algorithms to share content that confirms a user's beliefs and preferences. As a result, researchers have speculated that recommendation algorithms may unwittingly lead to the creation of online "echo chambers" or "filter bubbles" in which users engage content that increasingly intensifies their presumed beliefs. Although most of the research on "echo chambers" pertains to political partisanship rather than violent extremism, empirical support remains mixed, as echo chambers may be self-selected (i.e. users actively seeking to consume content carrying a certain message or viewpoint). Nonetheless, some scholars have theorized that online "echo chambers" could potentially play a role in the adoption of violent extremist beliefs and ideologies.
- **Gaming the algorithm:** While uncommon and some search engines protect against it, in theory users may attempt to "game" a recommendation algorithm by changing the content of their posts or reactions to them with the goal of prompting an algorithm to feature them more prominently. Possible methods include using specific keywords or hashtags, connecting to trending topics, encouraging an audience to interact with specific posts, and accessing the content via multiple

accounts. By way of example, these and related methods (e.g., "sticky-ing" and upvoting posts) were used by the moderators of "R/The_Donald" to dominate the front page of Reddit on several occasions before the tactic was banned.

- **Topic hijacking:** when a group of users leverages a trending hashtag to promote a topic that is substantially different from its original context. This most commonly occurs on Twitter and is relatively easy to carry out, since any user can include any hashtag in a tweet. If a group with enough users produces a sufficient number of tweets, it can arguably change the content and context of any hashtag. For example, the Islamic State hijacked hashtags during the World Cup in 2014 and again during the Black Lives Matter protests in 2020.

Ad tech algorithms

- **Product marketing:** Most ad tech content policies prohibit TVEC. However, where this content isn't moderated, extremist movements and terrorist groups can profit from their notoriety via ad tech algorithms that help to market specific clothing and merchandise. Many groups are associated with specific clothing items, while others are identified via slogans or memes that can be emblazoned onto commodity clothing items or merchandise. Terrorist and violent extremist groups may benefit from ad tech that increases sales of such clothing and merchandise in two ways: either directly, via revenue generation, or indirectly, in that ad tech repeats a particular message or raises the group's or movement's brand profile among consumer audiences.
- **Recruitment:** Although it is unlikely a group would directly recruit individuals within the ad itself, and many platforms prohibit ads or destinations that promote terrorism or violent extremism, in theory terrorist groups and violent extremist movements could potentially leverage ad tech to reach specific categories of individuals.

Table 1: Taxonomy of Content-Sharing Algorithms and Potential Exploits

Type	Examples	Potential Exploits
Search Algorithms	Web search Platform search Marketplace search	Data voids SEO Autocomplete
Recommendation Algorithms	Newsfeed Recommended videos, etc. Trending Topics	Echo Chambers Gaming Topic hijacking
Ad Tech Algorithms	First-party ad systems Third-party ad systems	Merchandising Recruitment

Note: given the similarity between recommendation and search algorithms, some exploits – such as data voids and echo chambers – can apply to multiple kinds of algorithms.

III. Literature Review

Whereas the section above mapped the theoretical pathways by which content-sharing algorithms might be exploited, this section highlights the key takeaways of a comprehensive literature review on the subject carried out by members of the WG. The review focused on articles between 2013 and 2021 that examined the relationship between content-sharing algorithms and engagement with materials related to terrorism, violent extremism, or radicalization. The review thus includes some studies that focused on TVEC material, as well as others that focused on content that may be harmful (such as hateful content that negatively targets out-groups, including political content).

Overall, the literature review looked at 11 studies that analyze the interactions between content-sharing algorithms and content related to terrorist and violent extremism. The studies can be found in Table 2 below, but key trends in the subject and research design of the studies include the following:

- **Platforms:** Platforms with more open APIs are more predominant in the literature, suggesting that researchers may be more likely to research platforms with greater ease of access. There is a prevalence towards research on YouTube; the platform's recommendation system was analyzed in six studies. Two researched Reddit and Twitter, and there was one each of Facebook and Gab.
- **Ideology:** Studies are weighted towards the far-right, which contrasted with terrorism studies more broadly, which tend to focus on Islamist extremism. Six

studies used far-right content exclusively, while four studied Islamist extremist content, and one utilized data from both ideologies. This may be because it has been easier to identify far-right content on social media platforms.

- **Language:** Data also tended to be English-language and focused on Western contexts. Five studies used English "seed" accounts; two used a mix of English and German; one was exclusively in German; two in Arabic; and one surveyed a range of languages (the seed account was predominantly made up of individuals from Southeast Asia and Iraq/Syria).
- **Research Design:** Not all the studies surveyed had content-sharing algorithms as their dependent variable. For six it was the primary focus of the study; two tested recommendation algorithms alongside other variables such as an "echo chamber" effect, while for one study content-sharing algorithms were the focus of one of the three research questions. Two others were wider research that offered findings on the topic.
- **Methods:** Only two studies created experimental conditions to account for user personalization and included baseline and control conditions. Four studies accessed YouTube's API to assess Related Videos – i.e., videos that are likely to be recommended (however, these cannot account for user personalization). One study drew from longitudinal Nielsen data to assess how users interacted online. Two studies accessed Reddit APIs to compare algorithmically sorted content against non-sorted, while another took this approach on Gab. Three studies took qualitative or observational approaches.
- **Time period:** most of the studies were published between 2018 and 2019. Two were published prior to the creation of the GIFCT in 2017. Only four were published after 2019 when YouTube and other platforms moved to restrict the visibility of "borderline" content.

In addition, key findings and takeaways from the studies include:

- Eight of the 11 studies suggest that content search and discovery algorithms can amplify extremist content. Four of these specifically looked at YouTube's recommendation system. One observational piece on Twitter showed that if a new account followed prominent Islamist extremist accounts, they would be recommended more. A study comparing a radical subreddit suggested that the most highly "upvoted" posts were substantially more extreme than a random sample. One piece of research found that Facebook's suggested friends had actively connected Islamist extremist sympathizers.
- One piece of research found no direct algorithmic effects but did find an interactive relationship between recommendation systems, echo chambers, and radicalization. On the other hand, another study found no algorithmic amplification of extreme content on either Reddit or Gab. In addition, two more studies suggested that YouTube's recommendation system either accounts for

little far-right content selection or actively discourages users from visiting extreme channels.

- Nine of the 11 studies focused on the **"supply"** of extremist materials – i.e., content with which a user could potentially engage – opposed to the **"demand"** side: discerning the media diets of terrorists/extremists. This imbalance is reflected in terrorism studies and creates a causal knowledge gap as to the role of content-sharing algorithms in online radicalization.
- Ten studies (perhaps unsurprisingly) focused **exclusively on the online domain** rather than considering how the online and offline milieus interplay. This runs the risk of inflating the existing policy concern around online radicalization creating a **"streetlight effect."** The one study that considers both domains notes that the two are inseparably intertwined.

It is also important to put these findings into their **wider context**. Many of the studies were carried out prior to the adoption of new tech company policies meant to limit the spread of content that may be related to extremism, such as YouTube's restrictions on borderline content. Little remains known about the effect of those changes.

It should also be noted that content-sharing algorithms do not exist in a vacuum and one should not overlook the extremist environment of which they are a part. In many instances, such as on Gab or the R/The_Donald subreddit, findings showed much of the content remained online because of lax moderation policies. In these situations, the null findings of algorithmic amplification are relevant because extremist content was readily available on the platform, suggesting that the users' own choice of platform may play a more important role than recommender systems.

The combination of the prevalence of one platform (YouTube) in the studies, combined with a focus on far-right ideology and Western-focused datasets identifies an important gap in the literature. Presently, little is known about how content-sharing algorithms have impacted terrorist movements in other regions around the world.

Table 2: Literature Review Materials

Study	Platform	Ideology	Language	Methods	Findings
Berger (2013)	Twitter	Islamist extremist	Arabic	Exploration of Twitter's recommendation system. Creates new account and follows Islamist extremist accounts.	"Who to follow" recommends a number of prominent Islamist extremist accounts.
O'Callaghan et al. (2015)	YouTube	Far-Right	English; German	Access API to draw Related Videos. Use text-metadata to categorize channels, which were checked against Freebase.	Recommends further far-right content that could result in "immersive ideological bubble."
Schmitt et al. (2018)	YouTube	Islamist extremist; Far-Right	English; German	Access API to collect Related Videos for two counter-messaging campaigns. Qualitatively analyzed and categorized 30% of dataset.	Extremist content within related videos. High crossover with anti-Islamist extremist campaign (possible due to keyword similarity).
Waters & Postings (2018)	Facebook	Islamist extremist	Multiple	Social network analysis	At least two ISIS supporters likely recommended as friends. Authors were also recommended IS-supporting accounts.
Ledwich & Zaitsev (2019)	YouTube	Far-Right	English***	Access API and use scraper to collect data on seed channels. Code into categories based on ideology and mainstream vs independent.	YouTube actively discourages users from extreme content. No evidence to suggest movement towards more extreme categories.
Ribeiro et al. (2019)	YouTube	Far-Right	English***	Audit seed channels that have been categorized into ideological groups. Access API to identify Related Videos and simulate navigation between channels.	YouTube recommends "Alt-Lite" and "Intellectual Dark Web" content, and once in these communities it is possible to find "Alt-Right" content, but not from recommendations. Suggest that findings support the notion of a "radicalization pipeline."

Reed et al. (2019)	YouTube; Reddit; Gab	Far-Right**	English	**YouTube/Reddit:** Create identical accounts, use bot to log in and engage with content. Access recommendations via API. **Gab:** Access data via API to compare "Recent," "Popular," and "Controversial" timelines. All: Code data according to Extremist Media Index (Holbrook 2015).	**YouTube:** Extreme and Fringe content more likely to be recommended and to be ranked higher. **Reddit/Gab:** Extreme material not promoted via recommendations.
Gaudette et al. (2020)	Reddit	Far-Right	English	Compare 1000 most "upvoted" posts in "r/The_Donald" against random sample.	Most upvoted sample substantially more extreme than random sample.
Baugut & Neumann (2020)	n/a*	Islamist extremist	German	44 interviews to explore media diet.	Individuals said that platform recommendations took them from basic knowledge to radical propaganda.
Hosseinmardi et al. (2020)	YouTube	Far-Right	English	Representative sample of web users' browser history over 4 years. Channels coded according to political ideology.	Pathways towards far-right content is diverse and only a fraction can be attributed to recommendations. No trend towards more extreme content over longer sessions. Suggest user preference plays a bigger role.
Wolfowicz et al. (2021)	Twitter	Islamist extremist	Arabic	Recruited 96 non-Twitter users. Treatment group suppresses algorithm, control group accepts all automated suggestions. Ask participants how they feel about suicide bombing.	Interaction effect between recommendations and network effects (i.e. filter bubble and echo chamber are complementary).

*** Data derived from interviews**

**** Includes male supremacism (i.e. incel/men going their own way)**

***** Not explicitly stated but examples or keywords are English-language**

IV. Knowledge Gaps

The preceding sections on mapping content-sharing algorithms and the literature review highlight several major gaps in our knowledge and understanding of the role content-sharing algorithms may play in violent extremist processes and pathways. The goal of this section is to categorize those knowledge gaps in more detail.

First gap: literature mainly focused on Western contexts and open platforms like YouTube

An immediate finding from the literature review is the paucity of research identified as directly applicable to the work-stream area of inquiry. A similar finding was an overarching focus of this body of research on the YouTube and Twitter platforms and content with a Western far-right ideology. The former is unsurprising given the ubiquity of YouTube and Twitter and the ability of researchers to compile and extract datasets from them. Yet it is problematic insofar as the impact of content-sharing algorithms on those platforms are taken as representative of their impact on all such platforms, and are no longer representative of the platforms themselves given changes they have made in recent years. As a result, the role content-sharing algorithms play in sharing content related to terrorism and violent extremism remains unclear.

In addition, there is also a tendency towards studies of far-right content on social media platforms. This may be a result of a number of factors, including the lack of Islamist extremist terrorist content on platforms due to takedown efforts, the ease of identifying far-right content, and the overall focus on Western-focused datasets. This combination of factors contributes to an overall lack of scope in present research as to how content-sharing algorithms operate in different global contexts.

Second gap: Transparent understanding of algorithms is still limited.

A less obvious but equally relevant finding is that while this body of research spans a wide range of data and methods, all but two of the studies adopt a methodological focus on algorithmic outputs (i.e., automated recommendations). The pros and cons of such an "output-only" methodological approach are well illustrated by one of the studies (Reed et al., 2019), which chose to co-opt the rigor of controlled experimental conditions to test the functioning of recommender systems across a number of platforms. The benefit of this approach is that it provides a more data-driven and empirically informed understanding of algorithmic outcomes. The limitation of such studies is that without any insight into how algorithms make recommendations, it is difficult to fully assess and understand how they may lead to different kinds of outcomes. Although several platforms have published blog posts that describe their algorithms at a high level and corporate engineers have occasionally published papers on recommendation systems, the information those sources provide is too general to offer insight into the potential

role they may play in sharing content related to terrorism and violent extremism.

Greater transparency into the role and functioning of content-sharing algorithms is therefore to be desired. However, legitimate questions arise as to what constitutes an appropriate level of transparency into how content-sharing algorithms are designed and developed that balances intellectual property rights, user privacy concerns, and the potential to share information that allows bad actors to exploit processes with the need to promote greater awareness and public trust. In this respect, emphasizing more transparent methods (such as evaluating an algorithm's inputs as well as outputs) as a means of demonstrating compliance with emergent voluntary and regulatory regimes may support such a balance. This in turn highlights how a more data-scientific approach to research, when combined with controlled access to industry design and testing methods, is essential to the production of more objective and verifiable research findings.

Third gap: very little literature on the role of human agency, or degree to which algorithms may exploit user behaviors vs bad actors consciously exploiting them

The third major gap concerns the paucity of literature on the role of human agency. There is tension as to whether algorithms are responsible for the exploitation of user behaviors, or whether bad actors consciously exploit them. There is also limited understanding of how people consume content on and offline. This is important as it influences where policy (both internal tech company and social media policy as well as external government regulation) should focus. To highlight the former, there is a need for more objective and verifiable findings to help inform collective understanding of the industry reality driving the development of such content-sharing algorithms, and often how the underlying "business model" is driving user/consumer metrics which are seen to be at fault. With respect to the latter, there is a need for more research on the specific affordances of various social media platforms to bad actors and what safeguards can be implemented to offset such vulnerabilities. To what degree the global public would benefit from further research into whether and to what extent such models may exploit behavioral tendencies or be exploited by bad actors is also a matter for consideration.

V. Future Considerations and Recommendations

Addressing the knowledge gaps identified above will be essential to better understand the potential role and impact of content-sharing algorithms (both positive and negative) on pathways to terrorism, violent extremism, and radicalization.

The following recommendations are offered to that end:

- **Widen the scope of scholarly research.** A comprehensive understanding of the role and impact of content-sharing algorithms will require both far more studies

of non-Western contexts as well as of platforms beyond YouTube and Twitter. Increasing the former will likely require new opportunities and incentives, such as research funding to provide researchers in non-Western contexts the means to conduct their own studies on how the dynamics of user-content and content-sharing algorithms operate in their specific contexts. Since some platforms are understudied by virtue of being private, the latter may also require new forms of privacy-preserving data access and novel multi-stakeholder collaborations (as discussed below).

- **Develop shared standards for measurement and evaluation.** Developing agreed-upon ways of measuring algorithmic outcomes will be essential to understanding their role and impact. Consider the knowledge gap above related to human agency. The extent to which malicious users may be actively manipulating content-sharing algorithms remains unknown, as does the extent to which content-sharing algorithms may be spreading content related to terrorism and violent extremism and influencing user behavior as a result. Yet the lack of understanding owes not just to insufficient data and research but to the lack of consensus about what data and research are needed to provide it. Making progress on understanding algorithmic outcomes will depend on developing shared standards around the kinds of data needed and how best to measure them.
- **Foster more data-sharing collaborations and public-private partnerships.** The main reason platforms like YouTube and Twitter are overrepresented in the research literature is that they provide public APIs that make open research possible, including measuring how different user behaviors lead to different content recommendations (and vice-versa). Although private platforms are not able to publish openly available APIs, carrying out research on those platforms should nonetheless be possible. For instance, with the appropriate safeguards, private platforms could grant authorized academics and researchers a way to sample from the outputs of an algorithm without having access to the underlying software or training data.2 Alternatively, they could find privacy-preserving ways of sharing data publicly, such as through the Social Science One dataset that Facebook released in partnership with leading universities and research centers. Regardless of the specific form such collaborations and partnerships might take, finding new ways for platforms to share data with the research community will be critical to improving our understanding of algorithmic outcomes.
- **Encourage more transparent explanations.** From the UK Online Safety Bill to the proposed EU Digital Services Act, policymakers have begun to focus on informing users (and thus researchers) about the reasoning behind algorithmic recommendations. Although such explanations will not in and of themselves be

2 In the technical literature, this was previously known as "black box" testing. `

sufficient to fully understand the role and impact of content-sharing algorithms, they will nonetheless make it possible to better understand the relationship between those algorithms and pathways related to terrorism and violent extremism (including potential "off-ramps" from radicalization).

Conclusion

While gaps remain in our collective understanding of the role content recommendation systems may play in amplifying TVEC across the online ecosystem, through this multi-stakeholder endeavor we have sought to take an agnostic approach to content type and map current industry approaches as a first step. With an increased understanding of such processes, practitioners may better identify opportunities to counter the online factors contributing to radicalization and thus better target the effective uses of positive interventions. In addition, further research could help inform policy discussion and development in this area as reflected in the recommendations.

The CAPPI WG recommends that GIFCT considers how to take this work forward through the next iteration of the GIFCT's WGs.

Acknowledgments

The multi-stakeholder participation in this WG has yielded significant benefits. Working groups are a multi-stakeholder effort to further discussion on the given topic of the nexus between terrorism and technology. This paper represents a diverse array of expertise and analysis coming from tech, government, and civil society participants. It is not a statement of policy, nor is this paper to be considered the official view of the multi-stakeholders who provided inputs. Special thanks to all the CAPPI working group members for their commitment to this multi-stakeholder group over the past 12 months.

Full List of Participating Individuals and Organizations

Aqaba Process	Microsoft
Brookings Institution, Chris Meserole (Co-Facilitator)	Mnemonic
Etidal	Moonshot
Facebook	Netsafe (NZ)
Global Partners Digital	New Zealand Government, Department of the Prime Minister and Cabinet (Co-Facilitator)
Google (Co-Facilitator)	Swansea University
Hope Not Hate	The Government of Ireland (Department of Justice)
Human Cognition	Twitter
Institute for Strategic Dialogue	United States Government (State Department and Department of Homeland Security)
Jihadoscope	Wahid Foundation
M&C Saatchi	Zinc Network
Maarif Institute	

APPENDIX

GIFCT CAPPI - LITERATURE REVIEW STUDIES

Baugut, P. and K. Neumann. "Online propaganda use during Islamist radicalization." Information Communication and Society 23, no. 1 (2020): 1570–1592. link.

Berger, J. M. "Zero Degrees of al Qaeda." Foreign Policy (August 14, 2013). link.

Gaudette, T., Ryan Scrivens, Garth Davies, and Richard Fink. "Upvoting Extremism: Collective identity formation and the extreme right on Reddit." New Media and Society (September 12, 2020). link.

Hosseinmardi, H., Amir Ghasemian, Aaron Clauset, David M. Rothschild, Markus Mobius, and Duncan J. Watts. "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube." (2020). link

Ledwich, M. and A. Zaitsev. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization." (2019). link.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems." Social Science Computer Review 33, no. 4 (2015): 459–478. link.

Reed, A. et al. (2019) "Radical Filter Bubbles: Social Media Personalization Algorithms and Extremist Content." Global Research Network on Terrorism and Technology 8. link.

Ribeiro, M. H., Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. "Auditing Radicalization Pathways on YouTube." Woodstock '18: ACM Symposium on Neural Gaze Detection. (2019). link.

Schmitt, J. B., Diana Rieger, Olivia Rutkowski, and Julian Ernst. "Counter-messages as prevention or promotion of extremism?! The potential role of YouTube Recommendation Algorithms." Journal of Communication 68, no. 4 (2018):758–779. link.

Waters, G. and R. Postings. Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook. Counter Extremism Project. (2018). link.

Wolfowicz, M., D. Weisburd, D. and B. Hasisi. "Examining the interactive effects of the filter bubble and the echo chamber on radicalization." Journal of Experimental Criminology. (May 2021). doi: 10.1007/s11292-021-09471-0.1.

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