



GDI

Global
Disinformation
Index

Disrupting Online Harms: **A New Approach**

Authors: Benjamin T Decker and Tim Boucher

Design: www.designbythink.co.za

The Global Disinformation Index is a UK-based not-for-profit that operates on the three principles of neutrality, independence and transparency. Our vision is a world in which we can trust what we see in the media. Our mission is to restore trust in the media by providing real-time automated risk ratings of the world's media sites through a Global Disinformation Index (GDI). The GDI is non-political. Our Advisory Panel consists of international experts in disinformation, indices and technology. For more information, visit www.disinformationindex.org



July 2021. Published under a Creative Commons License (CC BY-NC-SA 4.0)



Table of contents

GDI Christchurch documents collection: A retrospective	4
Disengaging hate: A call for new approaches to CVE	6
Radicalisation as a ‘marketing funnel’	10
Disrupting the funnel	13
Christchurch recommendations summary	15
Annexes	
Annex A: Example of harmful content working criteria (provisional)	19
Annex B: Example of harms comparison matrix (with criteria applied)	20
Annex C: Hybrid threat model	21
Annex D: Networked conflict dynamics	22
Annex E: Polarising and divisive content indicators	23
Annex F: Facebook and Twitter ad transparency	24

GDI Christchurch documents collection: A retrospective

In the two years since the Christchurch, New Zealand massacre, the countering violent extremism (CVE) community has been hard at work to find ways to disrupt the distribution of hate and extremist ideology online and at taking a retrospective approach by working to better understand the role the internet played in radicalising the Christchurch shooter.

Foreword

A recently published report from ICSR, [Far From Gone: The Evolution of Extremism in the First 100 Days of the Biden Administration](#), provides an overview of domestic violent extremism (DVE) in the United States—including groups and movements that flourished under the Trump administration and took part in the 6 January insurrection—and develops a new taxonomy on ideologically motivated violent extremism (IMVE). Among other findings, the report identifies the insurrection as the clearest example of how inefficient moderation policies failed to address the clear and present danger IMVE poses to the very core of democracy.

While the last few weeks have seen some very promising signs—namely, the United States [finally joining](#) the Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online—as well as the [publishing](#) of the Christchurch Call community Consultation report leading up to the [2nd anniversary of the Christchurch Call](#) (15 May), we still have a long way to go. As recently as this month, our team found a link to the livestream of the Christchurch massacre on 4chan, an indication that the video continues to circulate and can be found on the open web with minimal effort.

With these concerns continuing to linger, and with the goal of identifying the causes and warning signs leading up to these actions, we are increasingly referencing an internal white paper, the Christchurch Collection, produced by Ben Decker and Tim Boucher as part of our advisory role in the weeks leading up to the Christchurch Call to Action Summit in Paris on 15 May 2019. While the collection was written in the heat of the moment to help facilitate a global response to an unprecedented event, it continues to guide our approach to the issue of countering violent extremism (CVE) online. As you read through it, keep in mind that while the ideas in the paper are not necessarily original, the recommendations contained in the paper are intended to guide how we should move forward as invested parties.

From an editorial perspective, the original intent was to rewrite the paper in order to reflect how the landscape has changed over the past two years. However, when reviewing this collection, it became clear upon further contemplation that the speed of progress has been glacial. And as we enter this next phase of the pre-regulatory internet, the increase in case studies that have taken place—including those of the [Capitol attacker](#) who drove his vehicle into two police officers; the [shooting](#) of two police officers in Chicago, prior to which the suspect published multiple posts on Facebook announcing his intent to kill cops; and the insurrection on 6 January—prompted us to share the document collection as it was originally penned with the hopes that the proposed models can be contemplated, expanded upon and executed against by others in the field.

We've seen some signs of tech companies working together, or at least making similar efforts, to shift the tide. In a March 2021 [interview](#) with Stratechery, Microsoft's Brad Smith acknowledged that the Christchurch Call was a catalyst for the industry, requiring companies that comprise the 'technology stack' of the internet to align around public safety and stipulate which organizations should have what responsibilities when it comes to content moderation. Stemming from this, we've seen content moderation efforts and account bans/deplatforming, ranging from [figures included](#) in the [Disinformation Dozen](#) to QAnon content across [mainstream social media](#). While the short term has shown some iteration of success in this approach, we don't yet know the long-term impacts, nor whether they'll remain the best path forward.

However, what's missing from present-day conversations is the importance of counter-messaging strategies and what we call a 'Global Disengagement Centre' (pp. 4–5). The Centre, along with similar third parties, could help 'identify risky or problematic borderline content, including radicalisation materials, communities, actors, and adjacent content' and 'lead to taking proactive moderation and filtering actions against identified elements'. Deplatforming will likely always play a role in countering violent extremism; however, removing an amplifier from the equation doesn't stop ideologies from percolating—in fact, by the time a group or individual is removed, many of them have already been mainstreamed. Furthermore, the notion that removing 'influential' individuals fuels the 'censorship' narrative and Big Tech conspiracy theory has become prevalent—

a side effect potentially driving otherwise potentially reachable individuals further away from logic and into their polarised camps.

Take, for example, participants in the 6 January insurrection at the Capitol. While this is of course hypothetical, it's worth contemplating what the outcome would have been if more countermeasures had been taken to disrupt the radicalisation process and lessen the perceived sense of oppression and vulnerability felt by those who participated. A Washington Post [article](#) of 10 February helps quantify how many of these individuals felt, or were, disenfranchised in one way or another:

'Nearly 60 percent of the people facing charges related to the Capitol riot showed signs of prior money troubles, including bankruptcies, notices of eviction or foreclosure, bad debts, or unpaid taxes over the past two decades, according to a Washington Post analysis of public records for 125 defendants with sufficient information to detail their financial histories.'

What if, for example, an organization like [The Redirect Method](#) had been engaged, or a strategic plan developed with the list of stakeholders identified on page 4? The paper provides a multitude of tactics that could be deployed, including technical product interventions (page 14) as well the promotion of third-party counter-messaging campaigns (page 15). There is, of course, no way of knowing whether such efforts would have been successful. However, if we don't learn from the past and try new methods to prevent something similar from happening again, we may be doomed to repeat history.

As we publish this collection two years later, we wish to stress that we believe that the priority list of actions that were taken should have been different. Instead of a blanket ban, it is imperative that we explore ways to disrupt the radicalisation process and develop more robust and effective counter-messaging strategies. We look forward to thoughts, feedback and, of course, criticism—and most importantly, individuals and/or groups who are looking to forge partnership and help us develop a new path forward.

Kyle Orangio

Lead Analyst

The Global Disinformation Index

July 2021

Disengaging hate: A call for new approaches to CVE

Following the horrific atrocities committed in Christchurch, New Zealand in March 2019, the New Zealand Government has a unique opportunity to seek multilateral action in addressing previous gaps in web-based Countering Violent Extremism (CVE) efforts in both the public and private sectors.

Introduction

We recognise that CVE is a multi-pronged mission involving a variety of stakeholders in governance, technology, law enforcement and civil society. We adopt the Organization for Security and Cooperation in Europe's working definition of CVE as proactive efforts to: 1) counter efforts by violent extremists to radicalise, recruit and mobilise followers to engage in violent acts; and 2) address specific factors that facilitate and enable violent extremist recruitment and radicalisation to violence ([OSCE, 2018](#)).

As is evident from the amplification of violent propaganda after the shooting, **hate speech is platform-agnostic, sometimes bursting from the darkest corners of the internet to the most open public squares too quickly for any one company to intervene.** Addressing the connection between this content and the promotion of real-world violence thereby requires a deeper understanding of the size and scope of the technical and logistical challenges we face.

With the rise of the Islamic State and its global propaganda campaigns in recent years, both state and non-state actors collaborated on initiatives to eradicate violent extremist propaganda, including the [2017 Global Internet Forum to Counter Terrorism](#), as well as partnerships with civil society such as Jigsaw's [Redirect Method](#). Technology companies focused primarily on Islamic extremist actors, which, coupled with a multinational military operation in Syria and Iraq, **leaves many prior digital CVE initiatives' effectiveness unmeasured and their hypotheses unproven.**

Deplatforming and a host of technical tools featured in the content moderation space may have the ability to help us understand toxic hate speech at a macro level, but can they help us proactively make our cities and towns safer on a day-to-day basis? To what extent do digital tools like Natural Language Processing (NLP), Machine Learning (ML) and Artificial Intelligence (AI) play a role in a qualitative research battle where context reigns supreme over raw data?

As US Government CVE efforts failed to address the growing threat of far-right extremism both online and in the real world during the Trump administration ([GAO, April 2017](#)), the lack of global action in response to an increased convergence of the far right across social media platforms demonstrated a dangerous gap in our ability to maintain public safety ([Bellingcat, 2019](#)).

Moving forward, it is important to recognise the dynamic and platform-agnostic behaviours of the agents of disinformation and how they evade content moderation. We must treat this asymmetric power dynamic no differently than traditional pre-digital ecosystems, applying the same awareness of potential manipulation to social media and the networked actors ([Data & Society, 2018](#)).

It is with that in mind that we present the following recommendations, **urgently calling for the harmonisation of protocols across social media platforms, cloud service providers, and e-commerce providers.**

Top Policy Recommendations

1. Require public commitments from platforms to act on risk/harm minimisation principles.
2. Require platforms to take technical countermeasures to reduce the reach and impact of risky/harmful content associated with radicalisation to violence (diminish automated algorithmic amplification, and disrupt the radicalisation funnel).
3. Require platforms to engage with government, civil society, and other industry stakeholders to develop common standards for dealing with problematic content in real time.
4. Encourage platforms to promote third-party counter-messaging that directly challenges violent extremist narratives.

Recommendations

1. **Require companies to publicly commit to the principle of harm or risk minimisation.**
 - If a platform has the ability to prevent, reduce, or mitigate risky or harmful outcomes by taking corrective action, they have a duty to do so where the likelihood of harm by acting (e.g., removal) is lower than the probable or actual harm caused by inaction (e.g., allowing).
2. **Take technical measures to reduce the reach and impact of borderline content, using all available tools** (see section: *Disrupting the Funnel > Available tools*).
 - Disrupt the ‘radicalisation funnel’ at all levels through the use of all available platform tools.
 - Reduce the impact of the algorithmic amplification pipeline to prevent exposure to and consideration of radicalisation materials and communities in the first place.
3. **Enable threat-information sharing between companies and with governments** (‘ISAC’ model: *information sharing and clearing house* [Reference: [WaPo/Sen. Schatz](#)]).
 - Shared public standard for Objectionable Content across platforms.
4. **Promote third-party counter-messaging campaigns that directly challenge violent extremist narratives via a ‘Global Disengagement Center’** consisting of civil society stakeholders invested in CVE counter-narratives and real-world intervention strategies.¹
 - Fund, develop, and promote messaging campaigns through consultation with former extremists who have disengaged from violence-based extremism.
 - Promote confidential and secure intervention portals for CVE non-profit organisations to proactively reach affected individuals.
 - Incentivise companies to offer free content-marketing opportunities to promote SEO across all mainstream social platforms.
5. **Work with the ‘Global Disengagement Center’ and similar third parties to identify risky or problematic borderline content, including radicalisation materials, communities, actors, and adjacent content – and**
 - Take proactive moderation and filtering actions against identified elements.
6. **Harmonise protocols for defining a cohesive strategy to assess both individual initiatives and collective CVE efforts.**
 - Define measurable outcomes for monitoring and evaluating the effectiveness and impact of individual stakeholders against the commitments set forth in any global forum.

Definitions: Radicalisation to violence

Radicalisation: the process of exposure to, identification with, and indoctrination into a social narrative dividing the world into a righteous in-group and a less-than-human out-group; typically attributes threats to the out-group, justifies defence of in-group, and confers social value or status upon those who take radical, especially violent actions against out-group; spreads primarily through vectors of radicalised communities and materials (which induce participants into entering a kind of ‘marketing funnel’ – see below).

Radicalised communities: Groups of affiliated persons (especially online, in this context) who are at some stage of the radicalisation process, and who routinely interact with one another, often sharing, producing, or commenting on radicalisation materials.

Radicalisation materials: Inflammatory, divisive, and often hateful borderline content typically pertaining to social or political issues or actors, which is frequently associated with radicalised communities; contain one or several elements of the radicalisation narrative, function to indoctrinate, and act as a vector for further propagation.

Pre-radicalised population: Persons who have not yet been exposed to radicalisation materials or communities, and have thus not entered the radicalisation ‘funnel’; various populations have different levels of vulnerability to radicalisation depending on their perceived grievances and social status or stability.

Adjacent materials: Sensational, polarising, or outrageous borderline content which may itself not be clearly radicalising (i.e., may not obviously target an out-group), but which is often clustered or grouped with more inflammatory radicalisation materials, and may function as a radicalisation exposure vector, especially via the algorithmic amplification pipeline (see below).

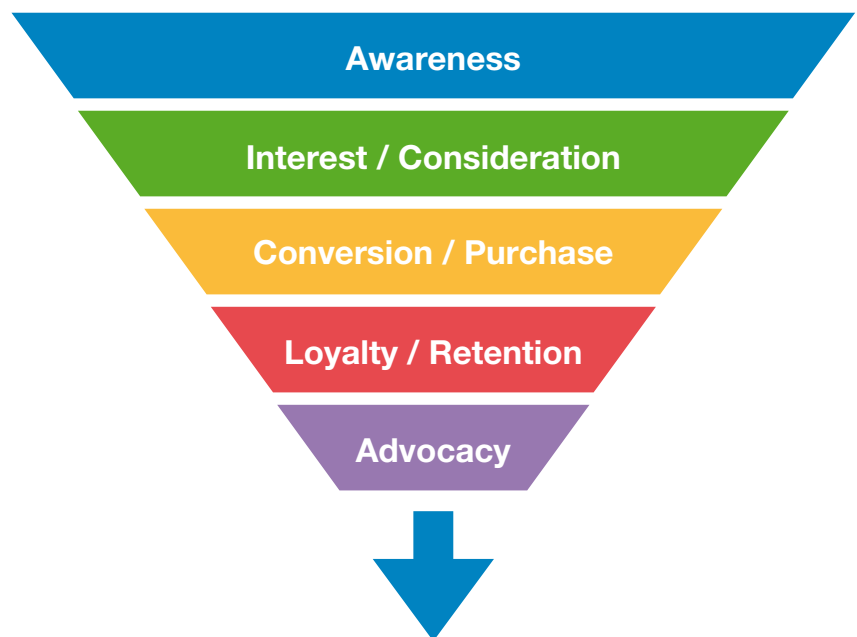
Borderline content: Content which comes close to violating the rules or community standards of a given platform, but which for some reason falls below the threshold of enforcement; content which is on the borderline of broad, or polite social acceptability (Reference: [YouTube](#)); may be adjacent materials, or potentially itself radicalisation materials.

Radicalisation as a 'marketing funnel'

Conventional models of radicalisation (such as the [Staircase Model](#)) focus on the stages of psychological transformation from unaffected or pre-radicalised persons through to the commission of violent extremist acts.

However, one common generalisation of the conventional marketing funnel includes the stages:

1. Awareness
2. Interest / Consideration
3. Conversion / Purchase
4. Loyalty / Retention
5. Advocacy



The 'customer journey' through the radicalisation funnel

Applied to the context of radicalisation, and married to key elements of the Staircase Model, the journey that a platform user takes through the funnel might look something like:

1. Pre-radicalised populations enter the top of the funnel, usually through **repeated exposure** to radicalisation materials and communities, especially via the 'Algorithmic Amplification Pipeline' (see below), which makes them consciously aware of the radicalisation 'product' or narrative.
2. Those who progress down to the middle of the funnel typically do so because their dawning awareness of the radicalisation narrative aligns with (mirrors) and activates (triggers) their own identity elements, including **beliefs, emotions, perceived grievances**, etc. Awareness turns into active interest and consideration as they discover and educate themselves via radicalisation materials and interactions with radicalised communities.
3. Unlike a conventional marketing funnel, there may not be any specific 'product' or 'service' which a potential 'customer' (i.e., radicalised person) must purchase to become a converted member of the radicalised community. Conversion, therefore, might be better framed in terms of **adoption (or identification)** rather than purchase. Adoption consists primarily of modelling behaviours and identity elements of perceived in-group peers within radicalised communities. Someone who has adopted the radicalisation 'product' themselves becomes a vector for transmission of new recruits into the top of the funnel.
4. When newly radicalised persons begin to adopt and outwardly communicate behaviours, identity elements, and radicalisation materials within relevant communities, they begin to receive **social rewards and status** within the new in-group. As these rewards/status may be something lacking in their ordinary life (making them vulnerable to targeting and exploitation), this peer-bonding also helps to solidify their indoctrination. During this stage, participants may also **reduce their social contacts** to persons who are not part of the in-group, or who don't reward them or recognise status.
5. Advocacy may begin nearer to the top of the funnel, when a new participant realises that the radicalisation product aligns with their interests, emotions, or beliefs. It increases drastically as they progress downward through the funnel. They also become more and more open to extreme versions of the underlying narrative, and calls to action by prominent community members. They may even issue their own calls to action, or begin planning or taking action themselves.

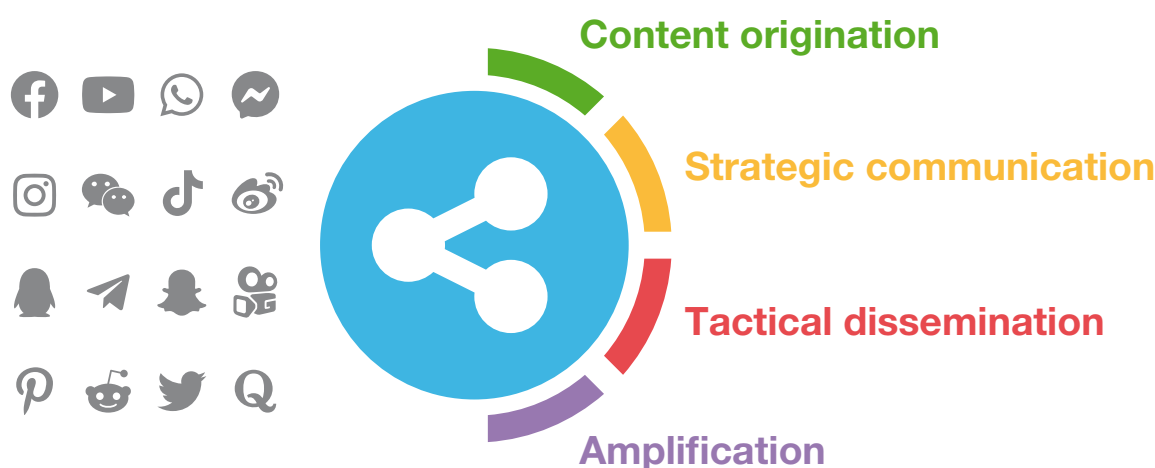
References:

- [The Staircase to Terrorism: A Psychological Exploration](#) (Fathali Moghaddam, 2005)
- [A Digital Funnel Drives People to Commit Hate Crimes in Real Life](#) (Quartz, Oct. '18)

The disinformation amplification pipeline

Social media algorithms are designed to amplify engaging content, often without regard to the nature of or risk associated with that content. (See: risk/harm minimisation principle above) Polarising and divisive borderline content can be highly engaging in that it generates strong reactions both for and against an issue. As engagement surpasses platform thresholds, its distribution may even be boosted automatically.

As a result, platforms automatically facilitate exposure of vulnerable, pre-radicalised populations to radicalisation materials and communities, pushing people to the top of the radicalisation funnel. Specifically tailored to user interests, recommendation algorithms (and auto-play functions, as in the case of YouTube) then automatically compel participants further down the funnel through the consideration/interest stage, and via high volume repetition may virtually guarantee conversion/adoption in audiences with compatible predispositions.



References:

- [How YouTube Built a Radicalization Machine for the Far-Right](#) (Daily Beast, Dec. '18)
- [YouTube's autoplay function is helping convert people into Flat Earthers](#) (Quartz, Nov. '18)

Disrupting the funnel

Ultimately, the utility of adopting the marketing funnel model for radicalisation is that it allows social media product features to be clearly mapped to stages of the radicalisation process, suggesting likely locations and methods for disruption and prevention.

Instead of attempting to de-radicalise affected populations (which may be a worthwhile and separate activity on its own), the goal herein proposed is **to actively disrupt all stages of the radicalisation funnel**. If radicalisation narratives are here compared to ‘products’, the desired outcome would be to increase ‘[churn](#)’—that is, to interrupt the ability for both pre-radicalised and already affected people to enter or progress through the funnel, by creating a poor, unwelcoming user experience.

The digital marketing funnel (how companies attract and retain customers via the web)



Available platform mitigation tools

It is recommended that all platforms retain the ability to choose the appropriate tolerances and tools to take action against radicalisation materials and communities. The list below represents a suite of common, currently available options, and different options may be most useful in disrupting different stages of the radicalisation funnel described above.

- Denial of service to known bad actors and radicalised communities (account suspensions)
- Identify, de-index, throttle, or remove radicalisation materials, communities, and actors from search functions, and other product surfaces involving distribution (such as automated recommendations)
- Reduce functionality of suspected, flagged, or automatically-identified risky borderline content (disable commenting, sharing, etc.)
- Create interstitial warnings (requiring opt-in) on borderline content
- Create age, phone number, credit card or other verification gates on borderline content (to reduce speed/volume of shares)
- Disable monetisation for borderline content (to reduce motivation for production/sharing)
- Disable API access or other automation/app integrations where problematic content is detected (to hedge against large-scale campaigns)
- Reduce or disable social rewards and status for sharing or producing borderline content (e.g., disable notifications, likes, RTs, reduce followers, etc.)
- Notify participants repeatedly of violations of platform rules or norms (negative reinforcement; to set clear expectations and create unwelcoming environment)
- Engage in and amplify active counter-messaging campaigns (run by third parties; e.g., '[Redirect Method](#)')

Christchurch recommendations summary

Necessity and urgency

- Preventing and countering violent extremism both online and offline according to applicable law is necessary for the maintenance of a free, safe, just, and democratic society.
- As violent acts shared online spur others to commit similar acts, there is a great urgency for platforms to take corrective action in stemming their spread.

Voluntary duty to not amplify

- Platforms should adopt in the short term a voluntary, self-imposed duty to not amplify harmful content related to violent extremism, and adjacent harmful borderline content, according to common, publicly available criteria.
- Platforms should collaborate with one another to develop voluntary industry standards and best practices for identifying and removing such content.

Establishment of a commission

- Governments should establish a commission consisting of partners from industry and civil-society groups to oversee the development of industry standards and definitions, with the intent of making recommendations for codifying them into law under a larger human-rights framework.
- The commission should first study existing tools and best practices, and work with platforms to measure effectiveness, iterate, and improve processes before making final recommendations about codifying them into law.

- The commission should investigate and identify safe harbour practices, including notice and takedown regimes for compliant platforms to follow, as well as fines, liabilities, and other appropriate disincentives and punishments where violations occur.

Development of a digital charter

- In the longer term, develop a larger online 'Terms of Service' or digital charter, which delineates the rights, privileges, responsibilities, and expectations associated with online communications integral to the perpetuation of a free and democratic society.
- It is recommended to codify only the most successful elements of voluntary industry standards and best practices, once they have been studied, road-tested, and improved.

Preliminary do-no-harm framework (pve/cve/crv)

- A preliminary do-no-harm (DNH) framework for social media PVE/CVE should be adopted, which would include publicly available working definitions to guide development and conversation around industry standards. (See: Annex A & B for examples)
- The DNH framework should be based on proximity to known violent acts, actors, and groups, and observable characteristics of the content, and applied as part of a risk-based analysis.
- The DNH framework should be tested in the field, iterated, and improved based on feedback, both from platforms who are implementing it and users who are affected by it.

Safe harbour for compliant platform participants

- Platforms that can demonstrate compliance with best practices related to the preliminary DNH framework should be rewarded with safe harbour where appropriate.

Error code for platform moderation and policy removals

- One means by which platforms can demonstrate compliance is by implementing a common error code (e.g., 'Error 452' for content moderation/policy removals) when they take actions against content hosted on their platform under the DNH framework and platform rules.
- Use of unique error codes for policy removals of content will facilitate transparency reporting.

Development of common notice and takedown regime for harmful content

- As abuse trends are constantly evolving, it is unlikely that any platform will be technically capable of detecting all instances of harmful content.
- User flagging and other third-party reporting should therefore be encouraged and facilitated.
- The commission should develop a common [notice and takedown](#) regime under which third parties can file complaints to platforms under the framework.
- Following the notice and takedown regime is another means by which platforms can demonstrate compliance and earn safe harbour.
- Any such protocol for notice and takedown must accommodate into its design mitigations against abusive uses.

Establishment of an Information Sharing and Analysis Organisation

- The commission should aid in the establishment with its partners of an information sharing and analysis organisation (ISAO) or centre (ISAC) to address emerging hybrid threats.
- Platforms should be encouraged to share hybrid threat-intelligence information under the framework with one another and with the government, in a way that respects applicable law, particularly human rights, including privacy.
- Participation would be another means for platforms to demonstrate compliance and earn safe harbour.

Hybrid threat model (summary)

- The threats posed by violent extremism and its online cultivation are of a hybrid type which exploits organisational vulnerabilities, definitional, and paradigmatic gaps, and which may be addressed at least in part by adapting cyber-security principles and techniques.
- Hybrid threats exhibit varying characteristics in novel and sometimes unpredictable configurations, and may include leaderless, loosely-organised, and spontaneous behaviours by actors of unknown number, location, origin, affiliation, or backing.
- Hybrid threats may be distributed across many platforms, actors, and accounts and may migrate or mutate in response to intervention efforts.

Intervention and disruption

- Interventions against harmful content under the framework should be proportionate to the risk associated with the material, actors, acts, and groups involved, and in accord with applicable law.
- Interventions should seek to disrupt the availability and spread of harmful content, while minimising the number of false positives.
- Taking into account the vast differences in resources between platforms of varying sizes, interventions should happen quickly enough that they can measurably stop or slow the propagation of harmful content in as short a time as is practicable.
- As de-radicalisation is a resource-intensive process that does not scale easily, interventions should be prioritised to prevent users from entering the top of the radicalisation funnel in the first place, and to disengage those in the middle and lower parts of the funnel.

Product intervention points

Platforms should take a nuanced approach to interventions, where different measures may be applied at different points in products depending on their nature.

- Intervention points include, primarily:
 1. the ability to post
 2. the availability of the post to the public
 3. the distribution and amplification of the post through the network
 4. advertising involving the post

Available technical product interventions and counter-measures

- Proactive filtering of known bad content types (includes hashes, keyword lists, domains)
- Warn the account-holder of violation and action required
- Remove the content: suspend the post and/or account

- Read-only mode (all posting is restricted)
- Reduce posting privileges related to suspect content
- De-index suspect content from search availability
- Throttle distribution of the content to prevent amplification
- Disable sharing, commenting, monetisation or other features on suspect content
- Display interstitial (opt-in) warnings on suspect content
- Require age, account, phone number, credit card, or other verification gates on suspect content
- Disable API/automation access to accounts engaged in amplifying suspect content
- Remove product reinforcement (notifications) and social rewards (likes, etc.) on suspect content
- Engage in and amplify active counter-messaging campaigns (e.g., '[Redirect Method](#)')

Promotion of third-party counter-messaging campaigns

- Directly challenge violent extremist narratives via a network of civil-society stakeholders invested in CVE counter-narratives and real-world intervention strategies
- Fund, develop, promote messaging campaigns through consultation with former extremists who have disengaged from violence-based extremism.
- Promote confidential and secure intervention portals for CVE non-profit organisations to proactively reach affected individuals.
- Incentivise companies to offer free content-marketing opportunities to promote SEO across all mainstream social platforms.

Timely restrictions on advertising involving harmful borderline content

- Reporting on the transparency of political and issue-based advertising, while a necessary piece of the puzzle, does not deter top-of-the-funnel radicalisation in real time.
- As advertising allows anyone to artificially boost the amplification of content (via paid placement), it is reasonable to request exchanges to rapidly implement a duty to not amplify harmful content (or domains associated with it) in the short term, whether in ad content itself, or in the user-generated content ads are run against.
- Advertising exchanges generally already have rules on allowed usage that are stricter than those of more open social media platforms, so further restricting certain ad content categories is in line with existing practice.
- Following the development of the commission's harmful content framework, companies should proactively monitor, identify, prevent, and remove all harmful borderline content from ads on their network.
- On ad exchanges which are also social media platforms (e.g., Twitter, Facebook), it may be prudent to consider whether accounts which post harmful borderline content in unpaid postings should be barred from access to political and issue-based advertising, and other ad buys.
- Given the increased vulnerability and legal protections due to minors, it may be prudent to consider limiting the ability of ad exchanges to target ads to anyone under 18 years of age, and to request the incorporation of an opt-out option for the targeting of advertising universally and for users across platforms.

(See: Annex F for examples)

Improve user verification procedures while respecting human rights

- The industry should collaborate on developing a more standard, open, and secure framework and tools for user-identity verification, which would be equally accessible to smaller companies with fewer resources
- Any such systems would need to be built according to a human-rights framework and be fully compliant with relevant international law (e.g., GDPR).
- While identity-verification of users may be beneficial in some ways, requiring real or verified identities should not be considered as a guarantee that infringing content or behaviours will not occur.
- Use of compliant identity-verification systems would be another means by which platforms could demonstrate compliance with the framework and earn safe harbour.

Annexes

Annex A: Example of harmful content working criteria (provisional)

The following could be considered higher-priority behaviours to capture under the provisional harms framework for PVE/CVE/CRV on social media, and are presented for discussion purposes only. They should be discussed, tested, and improved:

- Commission of violent criminal acts by account-holder (suspected)
- Depictions of actual violent criminal acts
- Incitement to violence
- Threats of violence
- Recruitment to join violent criminal group
- Depictions of violent criminal actors as heroic or inspirational
- Depictions of victims of violent acts which lack respect for the dignity or privacy of persons
- Trivialisation, encouragement, or celebration of suffering, violence, or death
- Promotion of the superiority of a group based on protected characteristics
- Promotion of discrimination, exclusion, or segregation based on protected characteristics
- Harassing, degrading, dehumanising language or behaviour
- Incitement to harassment
- Links to criminal violent acts, actors, or groups
- Links to harassing or hateful activity, actors, or groups
- Baseless, unsubstantiated, or defamatory allegations against persons or groups
- Not supported by facts; intends to or does disinform
- Raises likelihood of foreseeable, preventable actual harms

Annex B: Example of harms comparison matrix (with criteria applied)

Online Harms Matrix version 0.1 (PVE/CVE)	ISIS	CHRISTCHURCH	WHITE SUPREMACY	IDENTITARIAN	QANON/PIZZAGATE	HOLOCAUST DENIAL	ANTI-VACCINE	FLAT EARTH	AGENDA 21	ILLUMINATI	CHEMTRAILS	MOON LANDING	JFK ASSASSINATION	9/11 TRUTH	UFO DISCLOSURE	BIGFOOT	TOTALS
Commission of violent criminal acts (primary)	x	x	x	x	x												5
Depictions of actual violent criminal acts (social sharing)	x	x	x										x	x			5
Incitement to violence	x	x	x	x	x												5
Threats of violence	x	x	x	x	x												5
Recruitment to join violent criminal group or subculture	x	x	x	x	x												5
Depictions of violent criminal actors as heroic or inspirational	x	x	x		x												4
Depictions of victims of violent acts lacking respect for dignity or privacy of persons	x	x	x		x												4
Trivialization, encouragement, or celebration of suffering, violence, or death	x	x	x	x	x	x			x								7
Promotion of racial, ethnic, national, religious, etc. superiority	x	x	x	x		x											5
Promotion of discrimination, exclusion, or segregation	x	x	x	x		x											5
Promotion of fear, especially w/ sense of urgency	x	x	x	x	x		x		x	x	x		x	x	x		12
Named enemy, threat assignment to out-group, justifies defense	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		15
Calls to action (general)	x	x	x	x	x		x	x						x	x		9
Harassing, degrading, dehumanizing language or behavior	x	x	x	x	x	x											6
Incitement to harassment		x	x		x												3
Links to criminal violent acts, actors, or groups (secondary)	x	x	x	x	x												5
Links to harassing or hateful activity, actors, or groups	x	x	x	x	x	x		x									7
Baseless, unsubstantiated, defamatory allegations against persons or groups					x	x	x		x	x		x	x	x			8
Not supported by facts; intends to or does misinform					x	x	x	x	x	x	x	x	x	x	x	x	12
Used to sow doubt in face of evidence to the contrary					x	x	x	x	x		x	x	x	x			9
Promotion of distrust of government and rule of law	x	x		x	x	x	x	x	x	x	x	x	x	x			13
Promotion of distrust of social institutions and organizations	x	x		x	x	x	x	x	x	x	x	x	x	x			13
Promotion of distrust of science, including medicine							x	x			x	x					4
Raises likelihood of foreseeable, preventable actual harms	x	x	x	x	x	x	x	x	x								9
TOTALS	19	20	18	16	20	12	10	9	9	6	7	7	8	9	4	1	

Notes on the Comparison Matrix

- Categories of potentially harmful content have a binary rating (estimated) of whether or not a given harmful characteristic is observably and frequently expressed by material in that category. (This initial framework should be iterated and improved, with the addition of archived examples linked to definitions.)
- The items considered to bear greater potential or do actual harm are grouped toward the left, based on their proximity to violent criminal acts, especially, and to other common characteristics which are likely to violate existing platform rules, or to include related indicators associated with disinformation.
- The scores are simple tallies based on the presence of indicator criteria within a given category, and are presented only for the purpose of comparison. If improved and tied to specific content examples illustrating them, they could perhaps feed into overall category ratings.
- Individual artefacts within a given category could themselves be scored according to this or a similar rubric, and would inherit and feed into overall category ratings.

Annex C: Hybrid threat model

Background

Conventional approaches to detecting and countering online threat actors operating in social media and adjacent spaces tend to fall into one of two buckets at a national level:

- Foreign interference (intelligence agencies)
 - Hostile foreign powers attempt to influence or interfere with democratic processes and information within a target nation
- Domestic crime (law enforcement agencies)
 - Domestic actors commit acts which are against criminal or other law, requiring evidence, due process, enforcement by the legal system, etc.

Platforms apply their rules and available information to detect and take action (with varying degrees of success) against the violations including the above, especially using the model of 'coordinated inauthentic behaviour' (See: [FB](#), Dec. 2018)

Observations

Today's online threat landscape is significantly more complex and blended, which may lead to the majority of abusive and harmful behaviours slipping through the gaps between buckets or models.

- Example:
 - 'We're most concerned about fake accounts, because primarily that's the way bad actors try to do bad things on the platform.' – (Kevin Chan, FB, via [CBC](#), March, 2019)
 - 'Fake' accounts are only one part of a much larger ecosystem issue.

A more flexible model would more accurately reflect existing and evolving threats, and enable the mapping of prevention and mitigation solutions, as well as organisational competencies to address them.

Hybrid threat actor characteristics

Under the proposed model, threat actors would be considered 'hybrid' if they match several of the below characteristics (See: Networked Conflict Dynamics):

- Unconventional (falls outside past, known, or standard models)
- Emergent or spontaneous behaviour (including potentially unplanned/uncoordinated)
- One or several threat actors or operators, or unknown number
- Gradient of coordination, from little to none through to active planning and coordination of activities as a group on open and closed chat, forums, social media, etc.
- Blended authenticity: may combine authentic elements (identities, accounts, beliefs, grievances, real news stories, etc.) with inauthentic elements (fake accounts, satire, false news, etc.)
- May be distributed across multiple accounts and platforms, and where platform enforcement actions occur, may migrate readily to other platforms to continue
- Threat actors may themselves be domestic or international in origin (including from both overtly hostile and allegedly 'friendly' allied nations)
- Threat actors may receive a range of responses from tacit approval through to material support from states and quasi-state foreign powers (e.g., social media amplification), both wittingly and unwittingly

- Actual actions and content may or may not be illegal in the target nation (often non-illegal)
- Threat actors may move very rapidly, and may leave only ephemeral artifacts of short duration (e.g., platform suspension, or self-deletion), leading to difficulties for law enforcement to gather evidence and prove illegal acts, if present
- May operate just below the threshold of platform rules enforcement – usually intentionally
- Likely to pass undetected by platforms due to distribution across accounts (and across platforms), and, depending on the severity, may or may not be suspended as a rules violation (usually as coordinated inauthentic behaviour) if discovered
 - Unlikely to be reported to law enforcement or government if detected and suspended (unless duty of care is present, as in CSAM, etc.)
- May consist of a blend of both online and offline activities (e.g., an online campaign supported by rallies for and against an issue; violent or threatening offline acts)
- May make use of ads, social media posts with no placement cost, and peer-to-peer marketing (influencers)
- Relies heavily on viral propagation to induce secondary actors to re-transmit message
- Makes use of a blend of manual and automated production and distribution techniques

Annex D: Networked conflict dynamics

The landscape of today's disinformation conflicts were envisaged over a decade ago by visionary authors and thinkers such as Umair Haque, who wrote in the [Harvard Business Review](#) in 2008 regarding a paradigm known as fifth-generation warfare (5GW):

'4G war was network against state. Think Al-Qaeda vs America. 5G war is network against network, market against market, community against community.'

A 2009 [Wired](#) article by David Axe on fifth-generation wars offered additional insights which are chillingly accurate in retrospect:

'...[T]he next generation of war – the so-called “fifth-generation” – won't feature armies or clear ideas. It will be ... a “vortex of violence,” a free-for-all of surprise destruction motivated more by frustration than by any coherent plans for the future. 5GW is what happens when the world's disaffected direct their desperation at the most obvious symbol of everything they lack...'

Lastly, writing in 2006 ([Global Guerillas](#)), John Robb described characteristics of 5GW conflict actors as distributed, decentralised, spontaneous, bent on systems disruption, and using network effects to amplify disruptions. The correlations with observable traits of today's online bad-faith hybrid threat actors are clear.

Adversarial narratives

In the intervening years since these above predictions, we have unfortunately seen bad-faith actors of varying degrees of organisation actualising these concepts with troubling success. Hybrid-threat actors embody 5GW principles by combining the promulgation of adversarial narratives online with real-world violent extremist and terrorist acts. They then manipulate network effects on platforms to broadcast their message and recruit new members into their radicalisation funnel and continue the cycle.

Narrative analysis

Understanding and defending against adversarial narrative campaigns requires analysis of both the message contents and the context of online information artefacts and how they are propagated through networks. This analysis enables the identification of polarising and divisive indicators within content, which can be used to detect potentially risky artefacts matching established criteria to escalate their review by human analysts, and trigger proportionate interventions by platforms.

Annex E: Polarising and divisive content indicators

The following are common content indicators associated with harmful, borderline and radicalisation materials and communities online. Where multiple matching indicators are detected, it may indicate greater risk or likelihood of associated harms being present.

The indicators listed below are described in general terms, as they may be detected during the manual review of text content, especially, and to some degree may be automatically filtered and passed on to human reviewers (e.g., through custom keyword filtering lists and sentiment analysis). They are shared provisionally here with the intent of facilitating the further discussion and development of common definitions and sharable technical implementations.

Topics

- Social or political commentary, especially relating to issues of wealth, class, race, gender, orientation, as well as religious, political, and philosophical beliefs, identity, etc.

Tone and rhetoric

- Aggressive, outraged, or fearful in tone
- Hyperbolic, sensationalist, or exaggerated rhetoric
- Intended to elicit emotional response in reader

Psychological

- Emotionally triggering elements (identify-specific)
- Urgent call to action

Group identity

- Appeals to or rewards in-group identity, status, sense of belonging
- Coded language or unique references known only to in-group ('dog whistles')
- Reinforces existing beliefs of target audience (confirmation bias)
- Strongly differentiates between in- and out-groups

Escalation

- Assigns threat or directs hostility to out-group (names enemy)
- Justifies collective defence of in-group

- Moralises legitimacy of violence toward out-group
- Incitement to violence

Hate and dehumanisation

- Demeans, denigrates, dehumanises out-group or target
- Targets persons based on protected or quasi-protected group status or characteristics, including calls for exclusion (discrimination, hate speech)

Targeting

- Harassment or ad hominem attacks against a person or persons
- Discloses without consent confidential, sensitive, private, or a special category of protected data about a person or persons

Substance

- Alleges wrongdoing or criminal acts
- Victimisation claims (grievances)
- Attacks an organisation, or attempts to discredit an institution
- Posits nefarious control by invisible forces or actors
- Questionable historical claims, especially trivialising or downplaying suffering and atrocities
- Questionable medical claims, especially related to an elevated risk to persons or populations

Verification

- Novel and controversial claims without citations or verification by other sources
- Unsubstantiated and unfalsifiable claims (cannot be fact-checked, or impossible to disprove)
- Misrepresents known facts (can be fact-checked as false)
- False attribution, impersonation, and misrepresentation of sources

Context

- In-bound referrers containing many/most of above indicators
- Out-bound links containing many/most of above indicators
- Related social media conversations containing many/most of above indicators

Annex F: Facebook and Twitter ad transparency


Example of harmful content in Facebook ads

While Facebook's online Ad Library is a step in the right direction, it is plain to see that harmful borderline content associated with hateful and violent radical and extremist groups is not screened out.

It can also be seen in the below examples that ad buy amounts below \$100 are not carefully tracked, and that it would be inconsequential for bad actors to split up ad buys across accounts to stay below thresholds.

- Example FB ad search: "[Qanon](#)"
 - [Additional context: [SPLC](#), April 2019; [NBC News](#), Sept. 2018]

Ad Details



The Levitical Society
Sponsored • Paid for by Steve Guillotis

LOOK WHO WAS SEEN LEAVING THE SCENE OF THESE ONGOING CRIMES!!
KNOWN PERVERT PAUL REUBENS! CLEARLY JACKED UP ON
ADRENOCROME!! #Adrenochrome #PedoGate #PedoVore #PizzaGate
#TheGreatAwakening #KURU #TheStormIsFinallyHere #ThesePeopleAreSick #Q
#QAnon #FollowTheWhiteRabbit #WWG1WGA #opedohunt

NEW QAnon TRUTH BOMB!

www.facebook.com/TRUMPSTORM

CLINTON NECRO-PEDOPHILE CULT NOW IN BASEMENT OF ALAMO!

Data About This Ad

● Inactive
Aug 10, 2018 - Aug 11, 2018

<1K
Impressions

<\$100
Money spent (USD)

Who Was Shown This Ad

Age and Gender

Men Women Unknown

Age Group	Men	Women	Unknown
35-44	8%	0%	0%
45-54	21%	9%	0%
55-64	24%	11%	0%
65+	17%	11%	0%

Concerned Citizens Of America
Sponsored • Paid for by Concerned Citizens Of America

Click each hashtag one by one and see what they've been hiding from you

Tick tock #WWG1WGA #SethRich #NXIVM #Obamagate #ClintonCrimes
#Lolitaexpress #FollowTheWhiteRabbit #Bloodlines #SkullandBones
#EpsteinIsland #Podesta #Insurancepolicy #FastandFurious #JFK #MKULTRA
#Illuminati #Fisa #Reptilian #RoyalFamily #Adrenochrome #Spiritcor
#Bohemiangrove #Pedogate #Pizzagate #Mainstreammediacontrol #Vatican

Data About This Ad

● Inactive
Feb 26, 2019 - Feb 27, 2019

<1K Impressions

<\$100 Money spent (USD)

Who Was Shown This Ad
Age and Gender

Men Women Unknown

Age Group	Men	Women	Unknown
18-24	5%	0%	0%
25-34	5%	2%	0%
35-44	2%	2%	0%
45-54	7%	7%	0%
55-64	33%	7%	0%
65+	10%	12%	2%

Where This Ad Was Shown

Ad Details

Qanon Shall Win
Sponsored • Paid for by Qanon Shall Win

Donate today! We are going to as many rallies as possible to support our President Donald Trump and we need to make Qanon bumper stickers, signs, logos, decals! We need to spread the word and show our support!

<https://www.gofundme.com/qanon-rally-supply-drive>

Click here to support Qanon Rally Supply Drive organized by Qanon Shall Live

We will NOT let the fake media and the liberals think for even a moment that they can steal elections and get rid of our President Donald Trump! We have to

GOFUNDME.COM

Data About This Ad

● Inactive
Aug 11, 2018 - Aug 16, 2018

1K - 5K Impressions

<\$100 Money spent (USD)

Who Was Shown This Ad
Age and Gender

Men Women Unknown

Age Group	Men	Women	Unknown
35-44	3%	1%	0%
45-54	14%	5%	0%
55-64	21%	13%	0%
65+	19%	18%	0%



www.disinformationindex.org