CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

ASPI
AUSTRALIAN
STRATEGIC
POLICY
INSTITUTE

INTERNATIONAL
CYBER POLICY
CENTRE

PARTNERSHIP FOR COUNTERING INFLUENCE OPERATIONS

POLICY PERSPECTIVES SERIES #2

The Challenges of Countering Influence Operations

Elise Thomas, Natalie Thompson,
and Alicia Wanless

JUNE 2020

The Challenges of Countering Influence Operations

Elise Thomas, Natalie Thompson,
and Alicia Wanless

Carnegie's Partnership for Countering Influence Operations (PCIO) is grateful for funding provided by the William and Flora Hewlett Foundation, Craig Newmark Philanthropies, Facebook, Twitter, and WhatsApp. PCIO is wholly and solely responsible for the contents of its products, written or otherwise. We welcome conversations with new donors. All donations are subject to Carnegie's donor policy review. We do not allow donors prior approval of drafts, influence on selection of project participants, or any influence over the findings and recommendations of work they may support.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

CONTENTS

About the Partnership for Countering Influence Operations (PCIO)

Citizens, governments, and tech platforms around the world increasingly struggle to counter influence operations.

We believe that little progress will be made without a spirit of partnership between governments, the tech industry, media, academia, and civil society. Such collaborations are challenging but necessary in order to accomplish the three aims that PCIO believes are vital: to answer difficult policy problems related to influence operations; to find ways to understand the effect of adversarial influence operations; and to develop methods for measurement and evaluation of countermeasures.

PCIO is an international initiative, with partners and programming spanning multiple countries including in Latin America, Europe, and the Asia Pacific. PCIO and its advisory group will work actively to shape and promote an international, cross-sectoral consensus on key issues that is informed by evidence and best practice. PCIO leverages Carnegie's international networks, starting with its global centers, and is complemented by a select number of strategic partnerships. PCIO serves a convening function and as such does not speak on behalf of its members.

About the Policy Perspectives Working Paper Series

Influence operations cannot be solved by one actor alone, yet the field is ripe with mistrust and misunderstanding between industry and government. The PCIO's Policy Perspectives working paper series offers policymakers a primer on key issues in the field while helping to build consensus among stakeholders.

Summary

Influence operations are organized attempts to achieve a specific effect among a target audience. In such instances, a variety of actors—ranging from advertisers to activists to opportunists—employ a diverse set of tactics, techniques, and procedures to affect the decisionmaking, beliefs, and opinions of a target audience.

Yet much public discourse has failed to paint a nuanced picture of these activities. Media coverage of influence operations often tends to be negative, stoking fears about how influence operations might undermine the legitimacy of liberal democracies. But the purpose behind such campaigns must be considered when assessing their effects. In electoral settings, influence operations can refer not only to coordinated state efforts to influence a foreign election but also to positive advocacy campaigns such as those designed to encourage people to vote. That being said, governments and industry actors face growing pressure to do something about malign influence operations, but these campaigns must be clearly understood to be addressed effectively.

In reality, influence operations are neither inherently good nor bad, and it is up to societies themselves to decide what conduct and responses are and are not acceptable. The question of whether some influence operations are acceptable or not is highly ambiguous because it is so hard to ascertain the motives driving the actors behind them. This analysis examines a case study involving an online influence operation originating in Israel and targeting audiences in a host of countries including Australia, Canada, the United Kingdom, and the United States. The operators of this content mill push highly politicized online content and gain access to foreign audiences, at least in part for apparent financial gain.

This study explores the difficulties in making simple assessments about influence operations, and this particular case study serves as a basis for analyzing publicly available information about social media community standards and government legislation aimed at countering influence operations, for the purpose of identifying gaps in and challenges for the solutions proposed so far. This study ultimately points to the need for clearer consideration of what constitute acceptable persuasive techniques and tactics of online engagement, and it further highlights the lack of clear existing guidelines for policy development.

Key Takeaways

- **Influence operations defy easy categorization.** Influence operations often fail to fit neatly into boxes outlined by individual policies or legislation. They are run in a complex environment where actors overlap, borders are easily crossed and blurred, and motives are mixed—making enforcement challenging. In this case study, actors share highly politicized online content but also appear to benefit financially from their actions, making it difficult to ascertain whether their motives are primarily political, commercial, or both.

- **Relevant policies by social media platforms tend to be a patchwork of community standards that apply to individual activities of an influence campaign, not the operation as a whole.** Policies published by social media companies often focus on individual components of influence operations. This approach attempts to neatly categorize and distinguish actors (foreign versus domestic), motives (political influence and profit), activities (including misrepresentation, fraud, and spamming behavior), and content (such as misinformation, hate speech, and abuse). This piecemeal approach to enforcement raises questions about whether officials within social media platforms fully understand how influence operations work and how such campaigns are more than the individual behaviors that compose them.

- **Social media networks have more opportunities to counter influence operations through their platform policies than governments do with existing legislation.** Social media companies have implemented various policies to govern how their platforms are used, providing opportunities for combating influence operations. They also have greater access to information about how their platforms are used and have domain-specific expertise that allows them to create more tailored solutions. Fewer avenues exist for countering such influence operations using government-led legal mechanisms. This is not only because of the relative paucity of laws that govern online activity but also because law enforcement requires attribution before they can act, and such attribution can be difficult to ascertain in these cases. This means that governments have generally done little to help private industry actors determine what kinds of influence operations are unacceptable and should be combated. In the absence of such guidance, industry actors are de facto drawing those lines for society. Governments could do more to help guide industry players as they determine the boundaries of acceptable behavior by participating in multi-stakeholder efforts—some of which have been set up by think tanks and nonprofits—and by considering legal approaches that emphasize transparency rather than criminalization.

- **The influence operations uncovered by media scrutiny are not always as easy to counter as those writing about them might hope.** Savvy influence operators understand how to evade existing rules, so that their activities and content do not breach known policies or legislation. Media coverage that showcases examples of influence operations seldom explains whether and how these operators violate existing platform policies or legislation. This is a problem because distasteful influence operations do not always overtly violate existing policies or laws—raising questions about where the lines are (and should be) between what is tolerable and what is not, and, moreover, who should be determining those lines. Even when existing policies clearly do apply, these questions persist. Stakeholders should more clearly assess what constitutes problematic behavior before rushing to demand enforcement.

Introduction

Influence operations are organized attempts to achieve a specific effect among a target audience. Such operations encompass a variety of actors—ranging from advertisers to activists to opportunists—that employ a diverse set of tactics, techniques, and procedures to affect a target's decisionmaking, beliefs, and opinions. Actors engage in influence operations for a range of purposes. Covert political influence operations originating from foreign sources have been the subject of intense scrutiny in recent years and have stoked fears about how influence operations might undermine the legitimacy of liberal democracies. However, influence operations can also be motivated by commercial interests rather than conviction.

At times, there is no neat line dividing the two, a point that the case study featured in this analysis demonstrates. Churning out political content can be profitable, as the infamous Macedonian fake news industry, which captured headlines in the wake of the 2016 U.S. elections, demonstrated.[1] The more outrageous, divisive, or hyperbolic the content, the more clicks it drives and the more money website owners can earn from advertisers. The fact that this conduct is motivated by profit rather than political conviction does not reduce the potential political or social implications, including the possibility of inciting tensions or causing harm.

Some forms of influence operations (state-linked ones) are easier to handle. When state-linked covert information operations are discovered, they are likely to be in violation of online platforms' terms of service and therefore removed. Platforms like Facebook and Twitter have also taken to publicly disclosing their efforts to remove state-linked influence operations. The associated publicity imposes reputational costs on governments, political candidates, and other high-profile actors whose involvement in an influence campaign is discovered. These reputational costs *may* discourage them from trying again (at least in the same manner or immediately following public attribution of such ties).

More ambiguous forms of influence operations are harder to parse. Mixed-motive operations, in which operators benefit financially from sharing highly politicized content and the goal is therefore neither unambiguously commercial nor purely political, present a challenge for platform operators and regulators. When such campaigns are operated by private citizens with no official political affiliations, the costs of public disclosure may be much lower. Such an individual may not have a significant public reputation to maintain, and in some cases their activities may not technically violate a platform's terms of service without compelling evidence of particular motivations, which

can be difficult, if not impossible, to discern. And though media organizations are often quick to decry this behavior and call for enforcement, it is not always entirely clear if the behavior of mixed-motive operators is necessarily problematic. Some may argue that these actors have simply found clever ways to profit from the design of social media platforms, which reward highly engaging content.

Beyond this, the profit motive is a powerful one. Operators may have already sunk substantial time, efforts, and costs into establishing their so-called businesses. Strong incentives to continue are likely to make mixed-motive operations resilient, tenacious, and recidivist problems for social media platforms. This is especially true if the perpetrators suffer no real consequences after their operations are exposed (beyond, perhaps, some temporary disruptions as they replace the social media accounts they employ).

In either case, determining when an influence operation becomes problematic requires stakeholders to think holistically about the context in which an influence operation occurs, the actors who perpetrate it, the goals of their operations, the means by which they accomplish them, and the scale at which they operate. Influence operations can be problematic when operators attempt to disguise their identities or their aims, when they rely on false or misleading information, or when they cause real-world harm to their target audiences. In the absence of compelling ways to measure the effects of influence operations and their countermeasures and the lack of a whole-of-society approach to determining acceptable techniques of persuasion, influence operations continue to provoke much confusion and anxiety.

The case study examined here targets existing far-right, xenophobic, and anti-Muslim audiences on social media in an operation that appears to have both political and profit-driven elements. This case study highlights the way in which mixed-motive campaigns can fall through the cracks of existing regulatory frameworks, especially when they leverage preexisting and largely authentic social media audiences. The campaign centers on a set of thirteen websites that produce low-quality, inflammatory content that is then shared synchronously across a network of at least nineteen social media pages. The operators have gained moderator privileges on social media pages targeting audiences in Australia, Canada, the United Kingdom (UK), and the United States (among other countries) to share their content and drive traffic to the domains. Given characteristics shared by the websites, it is likely that they are run by a single operator or set of operators. Because the content produced by the domains and shared on social media platforms is often highly politicized and because the operators likely profit from advertising revenue from the domains, it is unclear whether they are motivated by political interests, commercial interests, or both.

This case demonstrates the limits of current ways of addressing this type of mixed-motive influence operation. Past reporting on this content mill identified it as part of a network of at least nineteen anti-Muslim Facebook pages pushing out synchronized posts,[2] resharing the same content within seconds of one another.[3] The posts tend to use Blogspot links to cloak the true sources of such content, such as freepressfront.com and speech-front.net.[4] The network was in part established by Israel-based Facebook accounts, which then approached users manning pages in other countries, offering to provide content to their audiences in exchange for also being made administrators on those pages.[5] This activity was linked by the *Guardian* to an individual named Ariel Elkaras in Israel; in the past, this individual had posted on search engine optimization and web marketing forums under the username Ariel1238a, seeking advice on how to monetize content.[6] When approached by journalists, Elkaras denied knowledge of or involvement in the content mill.

However, shortly after this exchange, several of the domains and a large amount of content were taken down.[7] (This study has not independently verified the attribution to Elkaras and will continue to refer to the group or individuals responsible for the content mill as the operators.) In response, Facebook appears to have removed a number of pages directly operated by the campaign, but it has not taken action against multiple other Facebook groups infiltrated by the content mill and its owners.[8] Many of the domains (some of which were briefly taken down following the *Guardian*'s 2019 reporting) are still active, and their content is still being widely shared across Facebook, Twitter, Reddit, and Gab. In short, the problem is far from over.

The Anatomy of an Online Content Mill

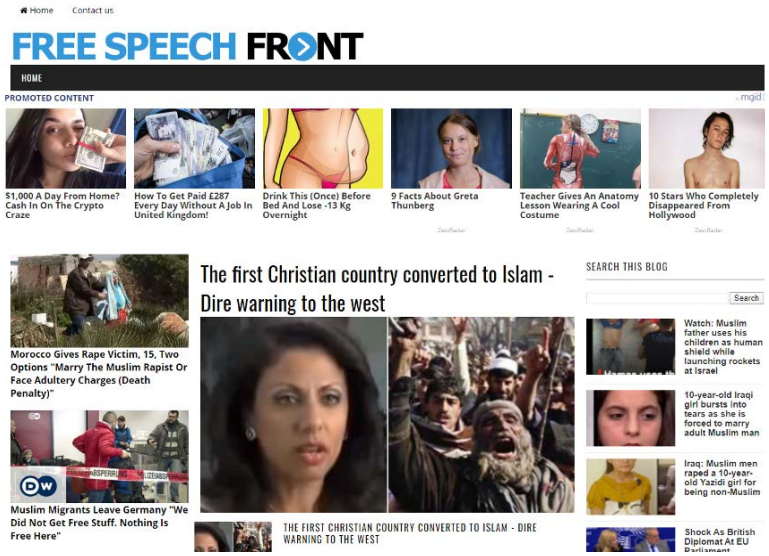
The business models of online content mills like the one featured in this case study rely on generating large amounts of low-quality web content (often either cheaply produced or simply plagiarized) to entice internet users to visit their websites, allowing the organizers of these sites to generate advertising revenue. The operators of content mills commonly use social media to drive traffic to their sites.

As of February 2020, the content mill in question still appears to be an active threat. Some of the domains it uses that were temporarily taken down appear to have been reinstated. And while Facebook removed the pages controlled directly by the content mill to promote its content, the campaign's operators are still active as administrators on a number of other far-right and anti-Islamic Facebook pages and continue to use them to churn out problematic online content. To understand how the content mill operates, it is vital to understand not only the content it produces and the domains it uses to host it but also the social media channels it employs to amplify its reach and attract readers.

The Content Mill's Content and Reach

Before delving deeper into the inner workings of this content mill, it is important to give a sense of the type of content it produces. The articles that the operators are publishing on the website domains they control have a relatively consistent format. Many of them are composed of between two and five paragraphs and relate to an embedded tweet or YouTube video from a range of sources. These low-quality articles consistently present Islam, in general, and Palestinian Muslims, in particular, in an overwhelmingly negative light.

SCREENSHOT 1.



The freespeechfront.net homepage

Most of the content is not necessarily overtly false, but it tends to be misleadingly slanted, cherry-picked, or otherwise taken out of context. Like much low-quality content online, the overall approach seems to be to present the most inflammatory narrative possible (see screenshot 1). And while there is little evidence to suggest that outright deception is the intention, the operators also do not seem to care whether the information they present is true.

For example, in September 2019, an article titled "Arab-Muslim Father Terrorizes his Infant Son With Gunshots" was posted on free-speechfront.info.[9] The article was based on a tweet from the Australian Jewish Association, commenting on a video of a man abusing an infant. The article reads:

> Arab-Muslim father decided to terrorize his baby son and share the video on social media. Instantly the video went viral and sparks outrage all over the world.
> The video posted on Twitter with the following description:
> "NEVER TOO YOUNG WHEN BORN FOR JIHAD! . . . Shocking video. Arab/Muslim father decides to acclimatise his newborn son to gunshots! . . . This is 'Palestinian' culture where hatred of Jews exceeds the love of children. Children are weapons of jihad, used as human shields and taught terrorism . . ."
> One of the top comments on the video referred to a famous statement by Israeli Prime Minister Golda Meir, who said: " . . . We can forgive the Arabs for killing our children. We cannot forgive them for forcing us to kill their children. We will only have peace with the Arabs when they love their children more than they hate us . . ."

The video in question was real but had no actual connection to Palestinians or Palestine. In reality, the incident took place in Saudi Arabia, and the perpetrator—believed to be the baby's brother—was arrested.[10] It is unclear whether the article's framing, which misplaces the video in the context of broader Arab-Israeli and Palestinian relations, is a case of accidental misinformation or intentional disinformation, or whether it was developed initially by the Australian Jewish Association or amplified by free-speechfront.info. The actual facts of the case would have been easy to establish, not least because multiple Twitter users replied to the initial tweet to point out that it was incorrect. What is clear is that tens of thousands of social media users were exposed to this misleading information. According to metrics from Buzzsumo, a proprietary research and analytics tool that provides information about the social media reach of online content, the free-speechfront.info article received over 39,900 engagements on Facebook, including 24,800 reactions, 7,300 shares, and 7,800 comments.[11]

Another illustrative example of the content mill's typical posts is an article posted on free-speechfront.net in November 2018, titled "Shock as Bulgaria, Romania, Serbia and Greece Declare War Against Radical Islam, the UN and the EU as They Join Forces With Israel."[12] The content of the article is a somewhat disjointed mash-up. The first half relates to Israeli President Benjamin Netanyahu's meeting at the Craiova Forum summit early that month with the leaders of the four countries, including his efforts to rally their support for Israel's positions in the United Nations (UN) and European Union (EU) on issues like Palestinian recognition and statehood.[13] (The article includes two embedded tweets from Netanyahu's official @IsraeliPM Twitter account.)

The second half of the piece is an attack on the UN and the abuse of Christian populations in Muslim-majority countries (including a paragraph lifted without attribution from a speech by Netanyahu on persecuted Christians in Iran).[14] This part of the article appears to have been copied from an earlier article on the since-deleted freespeechtime.net, which was part of the previous generation of domains connected to this particular content mill.

This example sheds light on the cheap, dubious ways content mills mass-produce and promote low-quality content. It demonstrates how content mills recycle stories—the aforementioned article uses Netanyahu's visit to the Craiova Forum as a news hook and then fills out the rest of the story with old and unrelated content. The same technique, and the same filler content, was used again for an article on politicsonline.net a few months later.[15] This article also demonstrates how content from the website is laundered through the broader information ecosystem. In addition to generating over 6,600 engagements on Facebook, 393 shares on Twitter, and hundreds of likes and retweets,[16] the article's content also spread across multiple other fringe and far-right blogs.[17]

Using the service Social Insider, this investigation identified the posts that received the highest engagement on Facebook pages infiltrated by the operators of the influence campaign (a tactic that will be discussed in more detail below).[18] Social Insider is a proprietary online tool that analyzes social media posts to identify which posts generate the most user engagement.

As an example, the most popular post (from February 14) on the "Guardians of Australia" Facebook page in the thirty days prior to February 24, 2020 linked to an article on thepolitics.online (see screenshot 2).[19] The article claimed, "A father from Afghanistan, barely making ends meet, sold his own daughter to a 55-year-old cleric for a goat and some food. Footage doing the rounds on the Internet shows the moment local women gave him a beating," and it included video of the Russian state-backed media conglomerate RT's coverage of a story from 2016. The opening paragraph of the article was found verbatim on numerous other websites including those of RT and former footballer turned conspiracy theorist David Icke.[20] At the time of the original RT story, the *Washington Post* ran an article on a similar subject stating that the girl had been rescued and was "in a shelter, while the man [had been] arrested and jailed," according to Afghan officials.[21]

SCREENSHOT 2.

The most popular "Guardians of Australia" post (February 2020)

The most popular post by a Facebook page called "Never Again Canada" for the same period featured an article about Auschwitz (as of May 2020, this Facebook page has been removed).[22] (The content mill often promotes pro-Israeli pieces in addition to overtly Islamophobic ones.) The group's second-most popular post promoted another thepolitics.online article containing a video titled "Watch: Adult Iranian Muslim Man Marries a[n] 11-Year-Old Girl."[23] The opening paragraph of this piece was also found on other websites such as trump-train.com, a so-called news site aimed at Americans but registered in India.[24]

The introduction was nearly the exact same as a Radio Free Europe/Radio Liberty (RFERL) article, which went on to note that the marriage had been annulled due to public backlash.[25] For its part, the article on thepolitics.online reported, "The girl is said to be around 11. The man is reportedly 33.

They were recently wed in a remote southwestern Iranian province with a video of the ceremony posted online." By comparison, RFERL stated, "The girl is said to be around 11. The man is reportedly twice her age. They were recently wed in a remote southwestern Iranian province with a video of the ceremony posted online." According to Google search return dates, both the RFERL and thepolitics.online pieces were published on September 4, 2019. This mishmash of directly copied or slightly reworded content taken from other sources is typical of the content published on the content mill's domains. This strategy is not unique to the operators—plenty of low-quality content online is derived from other sources—but the content mill does demonstrate consistency across domains in terms of the quality of the content shared.

Given that both of these examples are stories that were covered by more reputable media outlets, it appears that the content is not entirely false, per se. Rather, the operators tend to mislead by not specifying that both cases were resolved and the victims protected—follow-up information that was most likely available when the pieces were posted. Indeed, in the first example, over three years had lapsed between the Facebook post and the publication of the original article.

The Content Mill's Web Domains

Beyond the sort of content that the content mill seems to churn out, it is also worth examining the web domains on which the content tends to be published. The content mill operation is based on a series of domains dating back to at least 2017. There appear to have been several generations of domains, which were mostly Blogspot sites in the beginning but later became mostly independently registered domains. For example, the online content eventually migrated from sites like onlinepolitics.blogspot.com and freespeechtime.blogspot.co.il to ones like thepolitics.online and freespeech-time.com.

At least thirteen relevant domains were active as of February 10, 2020, and many of them appear to be interconnected. Ten of these domains use the same Google Analytics tracking code, which indicates that they are controlled by the same actor(s) (see table 1).[26] Google Analytics helps website owners track their web traffic. When website owners run Google Analytics on their sites, a unique eight-digit identification number linked to their Google Analytics account is inserted into the source code of the sites. One Google Analytics account can be used to track multiple sites. For example, one website could be marked UA-XXXXXXXX-1, with the repeated Xs representing the eight-digit identification number, while a second related website would be identified with the number UA-XXXXXXXX-2, and so on. These identification numbers can be used to discern which websites are likely run by the same set of operators.

TABLE 1

Suspected Active Domains Linked to the Content Mill

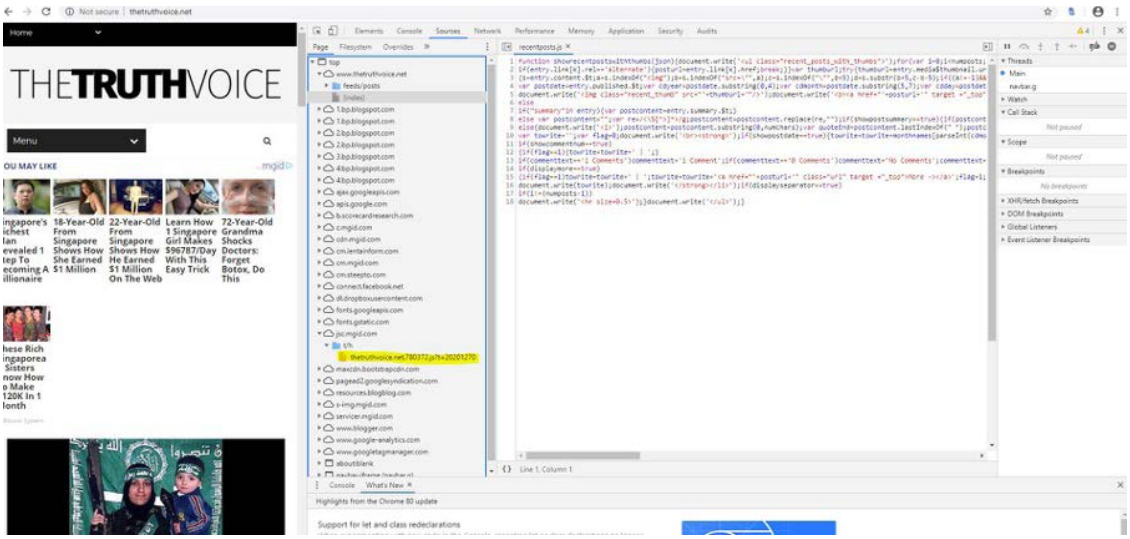
URL	Date Registered	IP	Archive	Google Analytics ID
thepolitics.online	1/27/2019	216.239.38.21	http://archive.is/Apvkr	UA-27810882-14
free-speechfront.info	1/14/2019	216.239.34.21	http://archive.ph/SUNtw	UA-27810882-11
freepressfront.com	12/22/2018	216.239.32.21	http://archive.ph/Ahm4F	UA-27810882-12
freespeechfront.net	1/14/2019	216.239.36.21	http://archive.ph/cYyOR	UA-27810882-9
i-supportisrael.com	6/7/2017	216.239.32.21	http://archive.ph/hApJC	UA-27810882-3
politicaldiscussion.net	4/9/2018	198.20.92.26	http://archive.ph/oS2Ot	UA-27810882-7
thetruthvoice.net	1/20/2019	192.64.119.249	http://archive.ph/ZZEls	UA-132814408-1
speech-point.net	2/6/2019	216.239.34.21	http://archive.is/RirNd	UA-27810882-10
speechline.net	9/21/2018	216.239.38.21	http://archive.ph/oS2Ot	UA-27810882-16
speech-point.com	10/16/2018	216.239.34.21	http://archive.ph/ygZUI	UA-157443622-2
threalnews.net	2/18/2019	216.239.32.21	http://archive.ph/LPA3J	UA-27810882-15
freespeech-time.com	11/19/2018	216.239.32.21	http://archive.ph/MzwZ4	UA-27810882-21
politicsonline.net	2/22/2019	216.239.32.21	http://archive.ph/MH0NH	UA-157443622-1

That is not all. An eleventh domain, thetruthvoice.net, uses a different Google Analytics code, but a close investigation of thepolitics.online (one of the initial ten linked domains) in February 2020 found that both websites share a different common identifying feature. At the time, they were both using the same MGid JavaScript, another type of analytics tracking script as seen below (see screenshots 3 and 4). This script appears to have since been removed from one of the domains (thepolitics.online).

Two additional web domains share a different Google Analytics tracking code from the initial ten domains but also appear highly likely to be a part of the content mill. The content from politicsonline.net and speech-point.com is highly consistent in tone and subject with the content shared by the other identified domains and often is shared by the same social media accounts. There is an obvious similarity between some of the domain names—between politicsonline.net and thepolitics.online, for instance, and between speech-point.com and speech-point.net; these striking parallels echo the similarities between other domains such as free-speechfront.info and

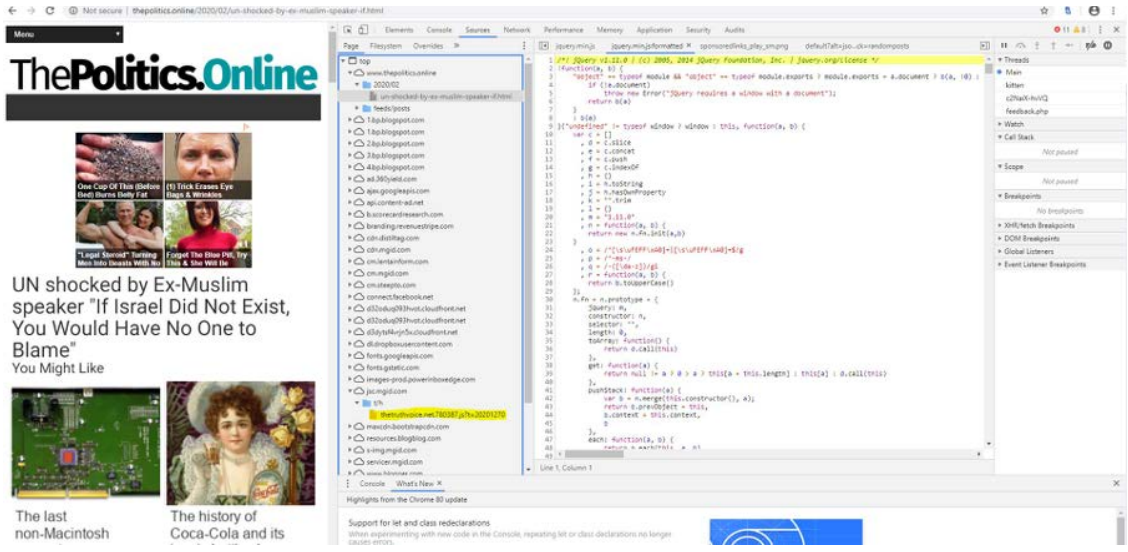
freespeechfront.net. The speech-point.net domain also uses the exact same wording in its cookie and privacy policy as five of the other domains.[27] Most obviously, speech-point.net, speech-point.com, speechline.net, free-speechfront.info, freespeechfront.net, and thepolitics.online all contain the sentence "First, Our blog [sic] designed to share legitimate political views while exercising our rights to freedom of expression and freedom of speech," and they all also include an invitation to contact the site owners at "the following email" without listing an email address.[28]

SCREENSHOT 3.



Use of JavaScript tracking code on thetruthvoice.net

SCREENSHOT 4.

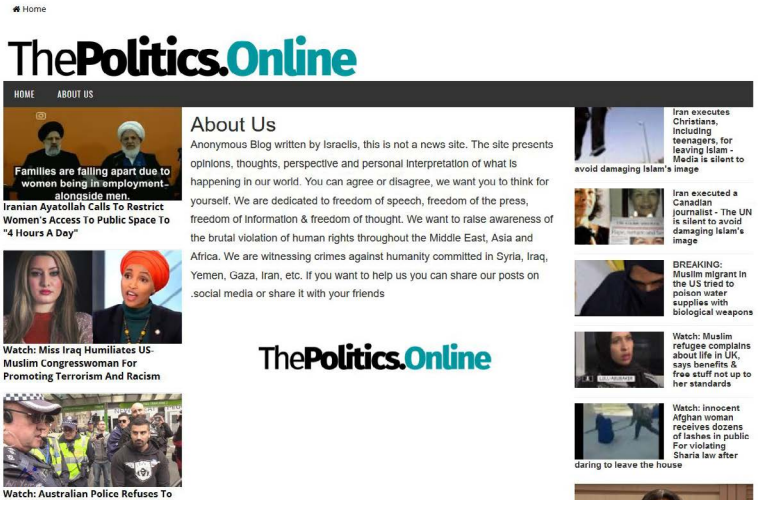


Use of JavaScript tracking code on thepolitics.online

There is further evidence of some overlap in content between the domains. For example, an article titled "U.S. Cuts All Funding and Announces Withdrawal From U.N.'s Cultural Agency Over Pro-Islamic Bias" was published on politicsonline.net on April 4, 2019.[29] The same article appears to have been previously published on freespeech-time.com in November 2018, although it has since been removed.[30] While these two domains have not been definitively proven to be part of the operation, it is possible that they are linked.

Most of these domains present themselves as news sites, with no overt information about who is responsible for running them. The exceptions to this are politicaldiscussion.net, a discussion forum, and thepolitics.online, which has an About Us page that describes the site as an "anonymous blog written by Israelis" and denies that it is a news site (see screenshot 5).[31] Despite this disclaimer, the site seems to be presented to look like a news site.

SCREENSHOT 5.



The "About Us" page of thepolitics.online

Overall, it appears plausible to conclude that these domains are run by the same actors. The similarity of the content in terms of style, tone, and subject matter as well as the repeated sharing of content across sites provide ample circumstantial evidence. The use of shared analytics accounts gives even stronger weight to the hypothesis that these domains are operated by the same individuals. These details demonstrate that, at the very least, eleven of these domains are connected through a single Google Analytics account, meaning that if there are multiple operators, they are working in coordination with the other sites.

For research purposes, the authors of this study used Uberlink to designate these domains as seed pages and conduct hyperlink network analysis of outbound and inbound links to these websites. Uberlink employs a web-based software known as VOSON to track and analyze when web pages link to one another through hyperlinks;[32] using the content mill domains as seeds, or points of origin, the software identifies all of the domains that link to the seeds or to other domains in the network.

In the case of the content mills' domains, this analysis returned a network of 143 domains with 239 connections between them. The visualization tool Gephi was subsequently used to create a network mapping of these websites, which is featured below (see figure 1). The mapping displays the relationships among the identified domains, all of which trace back (through some number of nodes) to the content mill domains.

This visualization reveals some surprising curiosities about the web domains that appear to be affiliated with the content mill. While the content mill has clearly targeted Western countries like Australia, Canada, the UK, and the United States, it does not target those countries exclusively. Nearly one-third of all the domains found in the hyperlink network analysis contained .in, the internet's top-level country code domain for India. Why would that be the case?

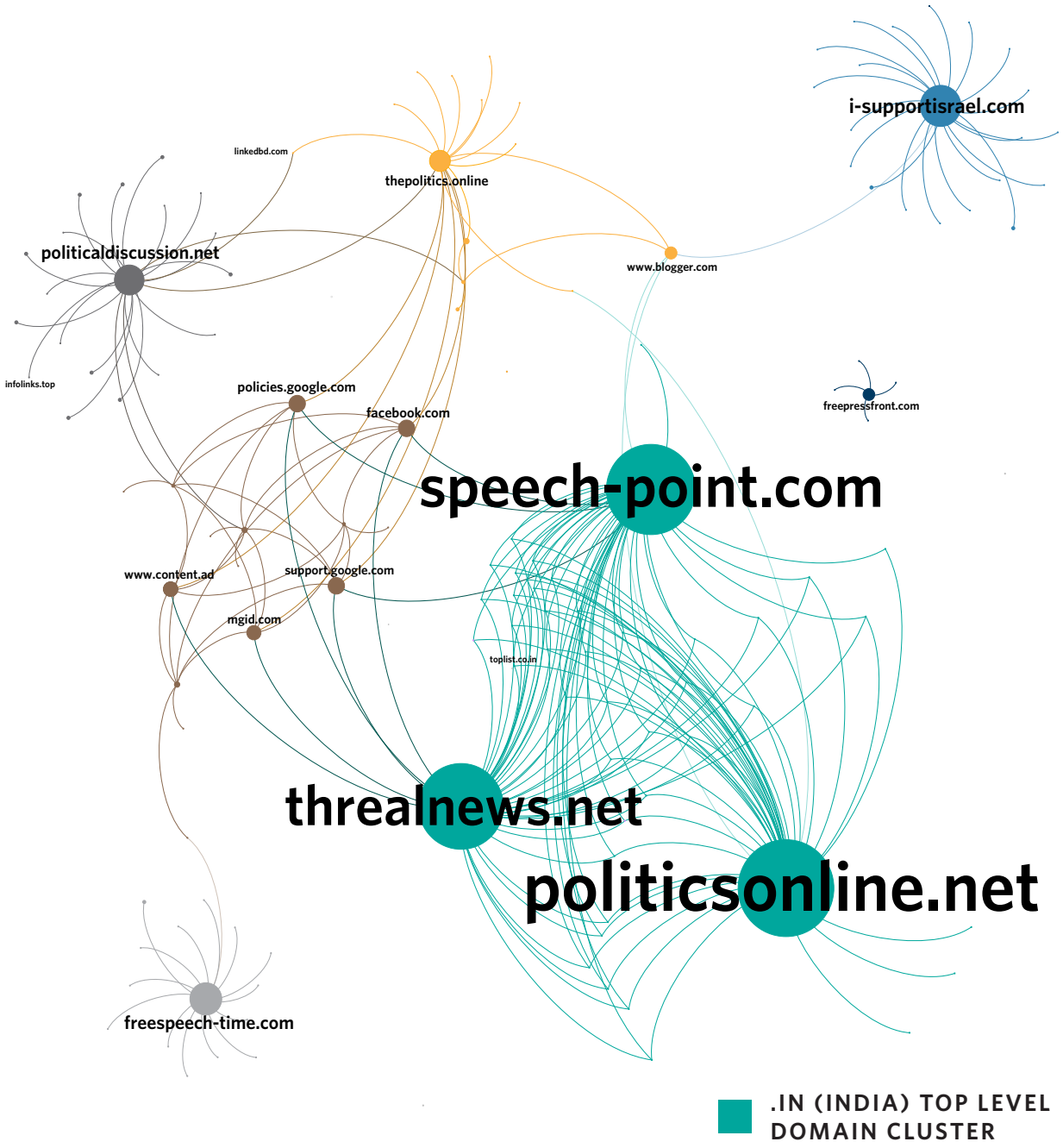
There are no definitive answers, but given the rise in sectarian violence targeting Muslims in India, it is possible that the operators see the country as a market for Islamophobic content or that India happens to be a serendipitous market for the purveyors of this inflammatory content.[33] Together with the websites speech-point.com, therealnews.net, and politicsonline.net, these websites formed a distinct community in the network marked by the color teal in the center of the network. Even though those India-centric domains form the biggest cluster in the network analysis visualization, this study focused on the content mill's activities in the West because the majority of the Facebook pages' content and domains in question were aimed primarily at American, British, and Canadian audiences, even if they were consumed by other audiences.

Another smaller community centered on the domain i-supportisrael.com, which is one of nine domains (6 percent of the total) that contained the term "i-supportisrael." A dozen domains (or 8 percent of the total) contained the term "Israel." This community is displayed in blue in the top right corner of figure 1.[34]

But the vast majority of domains in this network appear unrelated in nature to the content pushed by the seed websites. The primary purpose of these secondary domains may not be to spread content produced by the mill, but they are part of an ecosystem that helps boost such content in online search returns by increasing hyperlinks to them. Many of the connected domains appear to be directories listing and hyperlinking to other websites (such as www.toplist.co.in, www.linkedbd.com, and www.infolinks.top); this method of adding backlinks is one way to boost a website's search returns (and therefore advertising revenue).[35] This could further the claim that this operation at least partially is financially motivated rather than solely driven by political motives, but a group that wants to spread its content for political reasons might face similar motives to further its reach, underscoring the difficulty of determining the intent of online actors.

FIGURE 1

A Network Visualization of the Content Mill's Suspected Web Domains



NOTE: Nodes represent websites labeled by their domain name. Bigger nodes have more incoming and outgoing hyperlinks. The coloring represents communities of websites that link to each other.

The Content Mill's Facebook Page Infiltration

But the influence operation depends not only on the inflammatory content it creates and the domains that host it: it also relies on social media networks to help promote its work and attract readers. Internet archives show that the content mill previously operated its own Facebook pages, including pages linked to the free-speechtime.com and freespeechtime.net domains (in the latter case, the Facebook page was titled "We Love Israel").[36] The social media giant has since removed these pages.

Yet things did not stop there. Previous reporting by the *Guardian* uncovered evidence that the content mill operators have switched to using more covert tactics, infiltrating existing Facebook pages and using them to drive traffic to the content mill's domains. According to the *Guardian*, the operators of the campaign systematically contacted the moderators of established right-wing, pro-Israel, and anti-Islamic Facebook pages.[37] Posing as passionate volunteers, the operators offered to assist the moderators in operating these pages. On being given administrator privileges, the operators co-opted these pages to post content from their own domains.[38]

The content mill appears to have begun employing this strategy in 2016 using pages in the United States and Israel, before expanding over the course of 2018 and 2019 to include at least nineteen pages, according to the *Guardian*.[39] (BuzzFeed's reporting pointed to twenty-five pages.)[40] The campaign's operators are mainly targeting audiences in Australia, Canada, the UK, and the United States, though Austrian, Israeli, and Nigerian audiences have been targeted as well. To give a sense of how prolific the influence operation has been, the *Guardian*'s analysis found that the network posted 5,695 coordinated posts in October 2019, a month's worth of posts that generated 846,424 likes, shares, or comments over that period. According to reporting from December 2019, "the network [had] published at least 165,000 posts and attracted 14.3 million likes, shares or comments," according to the *Guardian* team.[41]

BuzzFeed has reported on the campaign's use of link redirection on some of the Facebook groups based in the United States and Canada.[42] Their investigation found multiple Facebook pages using links from Google's Blogger platform to disguise the real domain that hosts the content. For example, on Facebook, a link might appear to be from fpf-blog.blogspot.com, but when the user clicks, they are redirected to freepressfront.com instead. Experts that spoke to BuzzFeed suggested that this tactic might be an effort to evade detection by Facebook.[43]

This problematic behavior has not yet been stopped in its tracks. As mentioned previously, the BuzzFeed and *Guardian* investigations into the content mill were published in April and December 2019, respectively. As of February 2020, Facebook pages directly controlled by the operators appear

to have been taken down, but other groups that the operators have infiltrated are still running and still promoting their content. For instance, according to Facebook's transparency data (see screenshot 6), half of the eight administrators for the page "Guardians of Australia" are based in Israel. The page routinely shares content that has nothing to do with Australia, including posts from the content mill domains (see screenshot 7).[44]

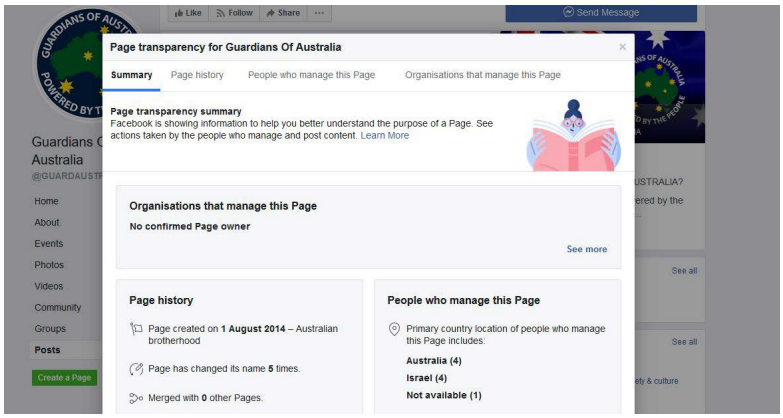
Interestingly, the surviving Facebook pages also often share content from Hananya Naftali, an influential Israeli social media figure and social media adviser to Netanyahu (see screenshot 8). While the content mill operators have a clear profit motive for driving traffic to their web domains, the intent behind their sharing of Naftali's content is more ambiguous. Such sharing could reflect a genuine desire to share the content, an effort to build a bigger audience for the pages by sharing popular content, or both.

The Content Mill's Other Social Media Outreach

While Facebook appears to be by far the most significant element of the content mill's social media operation, its operators do appear to be active on other platforms, including Twitter, Reddit, and Gab.

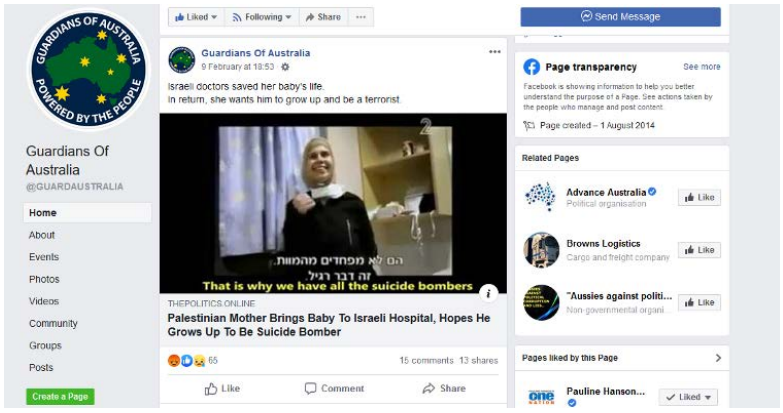
On Twitter, there are at least two accounts that appear likely to be a part of the content mill. One of these, @TimeSpeech, uses the

SCREENSHOT 6.



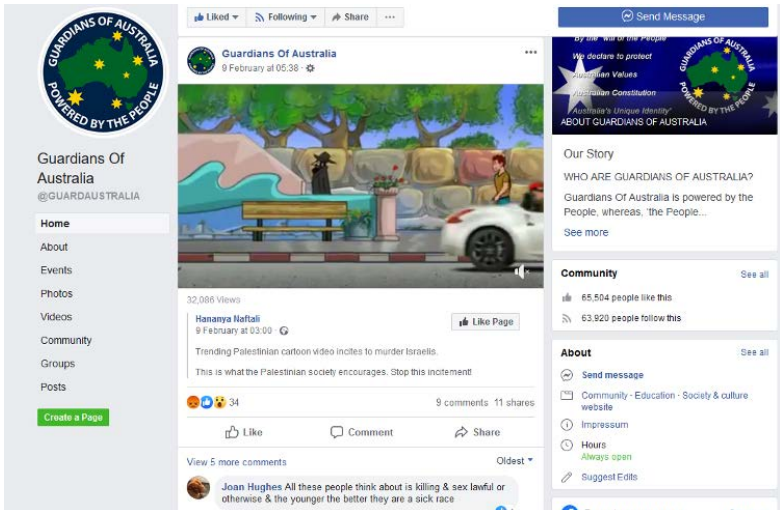
Transparency data for the "Guardians of Australia," Facebook page

SCREENSHOT 7.



Unrelated inflammatory content on the "Guardians of Australia" Facebook page

SCREENSHOT 8.

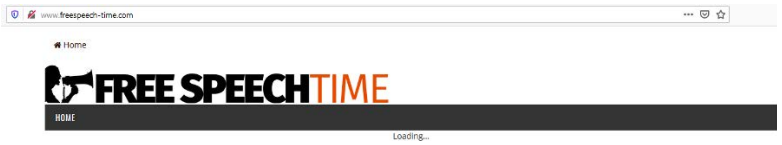


"Guardians of Australia" group's promotion of Hananya Naftali content

SCREENSHOT 9.

@TimeSpeech Twitter account

SCREENSHOT 10.



The website header of freespeech-time.com

same branding as freespeech-time.com and the group's now-removed Facebook page (see the screenshots 9 and 10).

The @TimeSpeech Twitter account shares content not only from freespeech-time.com but also from the other domains in the network (including speech-point.com and truthvoice.net). This reinforces the conclusion that they are likely part of the content mill even though they use different Google Analytics tracking codes).

Another domain, i-supportisrael.com, links to a Twitter account called @ISupport_Israel, which uses the same branding as the website. Both the @TimeSpeech and @ISupport_Israel Twitter accounts have pinned tweets promoting politicaldiscussion.net, a discussion forum that uses the same Google Analytics tracking code as the other content mill domains. The @ISupport_Israel account also shares content from all the content mill domains. Based on this evidence, it seems reasonable to conclude that these accounts are likely part of the content mill operation.

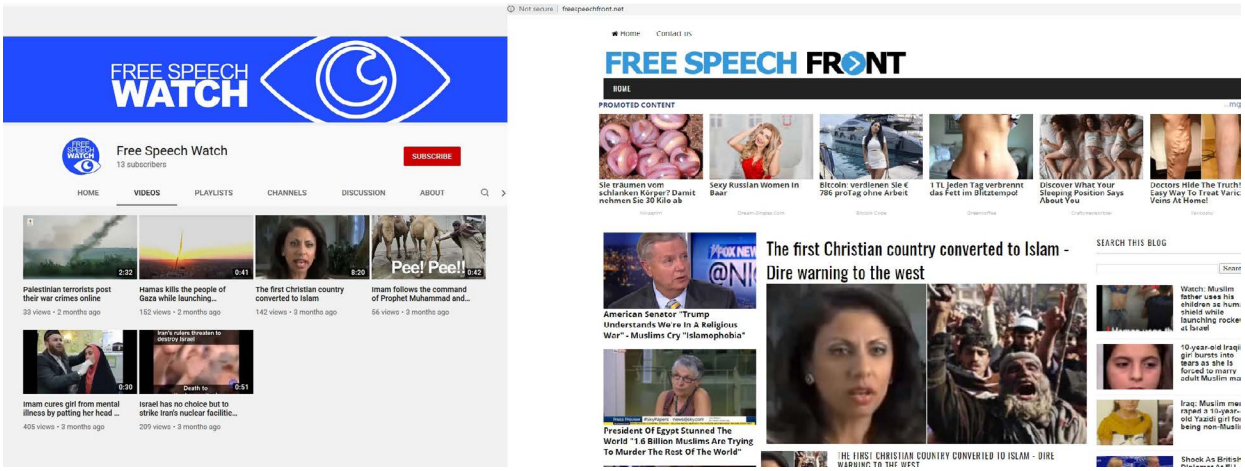
Three additional Twitter accounts are worth mentioning. All three use women's names—Sheila Berger (@SheilaB16315388), Alicia (@Alicia05972932), and Liza Rosen (@lizarosen101)—and have set a Star of David as their profile images. Most content that the three accounts share

comes from the content mill domains, and they sometimes retweet one another, as do the @TimeSpeech and @ISupport_Israel accounts. However, based solely on the simple facts that they often tweet content from the mill and retweet one another, there is not enough evidence to definitively conclude that these accounts are part of the content mill.

Meanwhile, four accounts on Reddit—unislamic, andynushil, Alisa1554, and lizalol151665—similarly appear to be dedicated to posting content almost entirely from the content mill domains.[45] They tend to post in conservative-leaning subreddits including r/Republican, r/The_Donald, r/HillaryForPrison, r/Conservative, r/IslamUnveiled, r/Australianpolitics, and r/exmuslim. The emphasis on U.S.- and Australia-focused right-wing subreddits echoes the influence campaign's choice of targeted Facebook pages.

The andynushil account is interesting for several reasons. It is the oldest of the four, dating back to 2017, and it is the only one with a verified email.[46] It uses the display name "FreeSpeechTime." The andynushil account broke character once in November 2019, posting in the Electronic Dance Music Production subreddit.[47] It is the only one to share videos from a YouTube account named "Free Speech Watch."[48] This YouTube account was created in November 2019 and appears likely to be the content mill's first venture onto the video-sharing platform (see screenshot 11).

SCREENSHOT 11.



Comparison of content on the Free Speech Watch YouTube account and freespeechfront.net

Interestingly, AndyNush is also the name of an account on Gab, a social media platform that markets itself as a place for "political speech protected by the First Amendment" and that has become a popular online home for far-right web users.[49] This account almost exclusively has shared content from the content mill domains.[50] At least one other Gab account, using the name Rachelrose,[51] has also been dedicated entirely to sharing the content mill's articles. The influence operation's apparent activity on Gab is a further indication of a tendency to target right-wing audiences.

In summary, the content mill is a persistent operation using at least a dozen or so domains and multiple social media platforms and accounts to spread inflammatory and misleading anti-Islamic and anti-Palestinian content. The evidence suggests a strategy of targeting right-wing and far-right audiences, but given that the operators stand to gain financially from driving traffic to their domains, they have clear commercial interests at stake as well. Despite multiple efforts by journalists to expose the group's activities, the content mill has not been significantly disrupted. Its content has been shared widely and continues to play on and contribute to existing social tensions.

Prospects for Public and Private Enforcement

What enforcement options do social media platforms and government actors have for addressing the kinds of influence operations that this content mill and others like it have employed?

Industry actors have put forth the most extensive options, yet even these measures are a patchwork of community standards that tend to apply to individual elements of an influence campaign without addressing the operation as a whole. Still, given the abundance of data available to social media platforms and their wealth of experience attempting to deal with malicious actors online, their policies are better tailored and specifically formulated to navigate the intricacies of online influence operations.

State-driven legal mechanisms for countering such influence operations are far less robust for a number of reasons: a paucity of laws that govern online activity and are well-suited to address influence operations, the difficulties of attributing malicious activity, and jurisdictional hurdles that may negate legal solutions even when perpetrators can be identified.

Platform Policies

Several problematic behaviors demonstrated by the operators of this influence campaign could be addressed by various existing Facebook, Twitter, YouTube, or Reddit policies. Those activities include:

- posting content that inflames social, political, or economic tensions;
- cloaking web addresses and surreptitiously redirecting online traffic to paid advertisements;
- using fake accounts to co-opt existing communities; and
- coordinating activities of ambiguous authenticity across platforms.

Reviewing the ways that social network platforms' existing policies could be used against the content mill operators could lead to a better understanding of the current opportunities, as well as gaps, for combating this type of influence operation. This analysis draws on publicly available information published by social networks and platforms. (Only platforms that this study found the content mill to be using were analyzed.)

The policies published by Facebook, Twitter, YouTube, and Reddit were similar enough to warrant combined analysis. The policies put forward by Gab, on the other hand, are extremely lax. Gab's terms of service are designed expressly to mirror the First Amendment protections of the U.S. Constitution, even though the amendment only restricts government actors, meaning that censorship by private parties like online platforms cannot actually violate or preserve such protections. As a result, while Facebook, Twitter, YouTube, and Reddit prohibit or reduce attempts to distribute or amplify hate speech, harassment, and false or misleading information, Gab allows such inflammatory and skewed content to flourish largely unchecked.

Gab's prohibited uses and content standards are minimal, going only slightly further than simply prohibiting conduct and content that are unlawful.[52] Gab prohibits obscenity, sexually explicit or pornographic content, certain commercial activities involving trade and financial instruments and sales involving animals or animal abuse, "unwanted advertising," and other unlawful acts, including unlawful threats or incitement to imminent lawless action. But its impersonation policies are loosely written to echo the broad protections of the First Amendment, making any kind of test of user authenticity or legitimacy seemingly impossible, and the policy on commercial spam "is not thought to encompass reasonable commercial use of a Gab feed by a commercial entity or individual to advertise your own commercial products."[53] These protections make Gab an ideal hub for extremist content that has been pushed off platforms like Facebook, Twitter, YouTube, and Reddit.[54]

Posting inflammatory content: The operators of content mills like the one featured in this case study appear to violate (in at least some instances) online platforms' policies on dangerous organizations, bullying and harassment, violence and incitement, and/or hate speech. Because of this, these policies are one potential avenue for countering the online actions of such parties. Yet such policies can be challenging to enforce, as many savvy influence operators know how to create content that falls below the threshold of enforcement but is nonetheless harmful.

Apparent ties between the content mill's operators and extremist organizations could be grounds for enforcement, but proving such connections conclusively is no easy feat. The team of reporters from BuzzFeed and the *Toronto Star* that first reported on this influence campaign pinpointed ties between the operation and the "Never Again Canada" Facebook page, which had ties to a noted right-wing extremist organization known as the Jewish Defense League.[55] These reporters highlighted that

Facebook had previously removed this extremist organization from its platform citing its dangerous organizations policy. Twitter and YouTube similarly prohibit violent extremist content, meaning that any overt ties to extremist groups would be enforceable across all three platforms.[56] In the case of this content mill, however, evidence of ties to extremist organizations appears to be mostly circumstantial, making many of these policies on dangerous organizations largely inapplicable in this instance.

Beyond circumstantial ties to an extremist organization, the operators' online posts could potentially run afoul of platforms' policies against harassment, the incitement of violence, and hate speech, though enforcement of those standards has been uneven. The posts the operators have shared on Facebook, Twitter, and Reddit appear to have routinely expressed anti-Muslim sentiments that could be subject to policies concerning bullying and harassment, violence and incitement, or hate speech. It is worth noting that these platforms enforce policies on restricting hate speech, violence, and incitement even when such words are not targeted at a specific individual—the wording of all four of the platforms' policies makes explicit reference to content that targets groups of people based on protected classes. These protected classes or characteristics vary between platforms but typically include race, ethnicity, national origin, caste, sex/gender, religion, gender identity, sexual orientation, and disability. Other broader identifying markers are sometimes included too, such as age, diagnosis of a serious disease, veteran status, immigration status, and status as a victim of a violent event or such a victim's kin.[57]

It is worth looking a bit more closely at each specific company's policies. Facebook's bullying and harassment policy covers "unwanted malicious contact" and other harms that fall short of violent threats or hate speech.[58] The social media giant's policy on violence and incitement covers statements that aspire to, condone, or intend violence, and Twitter's policy covers statements of "intent to kill or inflict serious physical harm on a specific person or group of people."[59]

Facebook's hate speech policy categorizes various tiers of harmful content. The most egregious content contains violent or dehumanizing speech and targets individuals or groups based on protected characteristics or immigration status.[60] Second-tier content involves generalities or statements that express a group's inferiority or contain targeted cursing.[61] The third tier, perhaps the one most applicable to the content posted directly by the pages associated with this influence campaign, includes calls for political, economic, or social exclusion on the basis of protected characteristics, including "race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and diagnosis of a serious disease or disability."[62] Facebook policy prohibits all three types of hate speech.

Meanwhile, Twitter's hateful conduct policy forbids individuals from promoting violence against, threatening, attacking, or inciting fear on the basis of a protected category, as well as using language that dehumanizes individuals on the basis of religion.[63] And their policy on abusive behavior forbids users from wishing serious harm on others and using "aggressive insults with the purpose of harassing or intimidating others."[64] YouTube platform policy similarly prohibits "content promoting violence or hatred against individuals or groups" based on protected categories.[65] Reddit's policy against threats, harassment, and bullying is quite expansive and includes "anything that works to shut someone out of the conversation through intimidation or abuse, online or off[line]," and includes menacing or directing abuse at an individual.[66] It is worth noting that platforms enforce policies against hate speech, violence, and incitement even when such words are not targeted at a specific individual—all four of the platforms' language makes explicit reference to content that targets groups of people based on protected classes.

The content promoted by the content mill in question is often inflammatory, offensive, and deeply skewed. Yet, absent a specific threat, call to violence, or element of dehumanization, the content can often evade detection and enforcement under online platforms' policies. The *Guardian* found that the operators behind this campaign targeted violence against Australia's first female Muslim senator, Mehreen Faruqi, as individual followers launched a torrent of aggressive comments against the politician on Facebook.[67] Incidents like these would likely warrant enforcement under Facebook's policies on hate speech, violence and incitement, or harassment, and the other platforms have similar prohibitions on targeted harassment.[68]

Whether or not the policies more broadly apply to the bulk of the content visible in the archived links is somewhat debatable, as the posts use false or misleading information to suggest negative stereotypes or characteristics but fall short of outright dehumanizing or hateful language. That said, as the *Guardian* report mentioning Faruqi put it, "the posts stoke deep hatred of Islam across the western world and influence politics in Australia, Canada, the UK and the US by amplifying far-right parties."[69] Given the ambiguity of some of the policies mentioned above, perhaps there is a case for arguing that the content in question creates a pervasive culture of hostility toward a protected group that could justify enforcement.

While platforms can escalate enforcement options in response to repeated offenses—Twitter, for example, suspends accounts that repeatedly violate its platform rules[70]—this consequence occurs in cases where clear violations have taken place. There is little to suggest that existing platform policies are meant to address cumulative harms. Enforcement on these grounds would likely require a rein-

terpretation of existing rules and guidelines, which largely focus on individual instances of harmful content or coordinated *behavior*, regardless of content. This may be easier for platforms like Reddit, for example, whose policies are vaguer than those of Facebook and Twitter, as Reddit may have more latitude to interpret the rules according to specific harms.

Cloaking URLs and redirecting traffic to paid advertisements: Operators of this content mill and others like it also may be violating online platforms' terms of use by disguising the web addresses (also known as uniform resource locators or URLs) of their content to covertly siphon off web traffic and boost their advertising revenues.

Facebook, Twitter, and Reddit explicitly prohibit the practice of URL cloaking and redirecting users to websites that appear irrelevant to the posted content. Facebook outlines this stance under its spam policy, labeling it both "cloaking" (the bait-and-switch tactic of showing one link and then redirecting users to a different site) and "misleading content" (showing a link that promises a certain kind of content but delivering another kind instead).[71] Reddit also bars "posting content that includes link redirects as a way to circumvent an existing domain block and/or disguise a link's source."[72] Twitter's policy explicitly prohibits commercially motivated spam, which would appear to be the closest approximation of the behavior exhibited by the operators in this case study and should easily apply to the offense in question.[73]

Using fake accounts to co-opt existing online communities: When the operators of this influence campaign reached out to existing online communities and requested administrator access to their Facebook pages and groups, it appears that they may have done so using inauthentic Facebook accounts.

Facebook has two existing policies that could prevent a threat actor from using fake accounts to gain editor privileges on popular pages: its misrepresentation policy and its inauthentic behavior policy. The misrepresentation policy bans users from creating inauthentic profiles and prohibits individuals from recreating profiles that have been taken off the platform.[74] The inauthentic behavior policy further prohibits users from "mislead[ing] people or Facebook about the (1) identity, purpose, or origin of the entity that they represent; (2) popularity of Facebook or Instagram content or assets; (3) purpose of an audience or community; [or] (4) source or origin of content."[75]

The other social media platforms—Twitter, YouTube, and Reddit—have similar prohibitions on inauthentic or misleading activity, but they vary in stringency. Because Twitter, YouTube, and Reddit do not explicitly require users to share their true identities, users are free to potentially have multiple accounts. Problematic behavior arises when these multiple accounts claim to be representing individuals or organizations they do not actually speak for or when they are used to engage in harmful behavior or artificially generate engagement on the platform. Twitter refers to such tactics in its

policy on "platform manipulation," writing that the use of multiple or fake accounts to "engage in spamming, abusive, or disruptive behavior" or to "artificially influence conversation" is prohibited.[76] Similarly, YouTube forbids individuals and channels from impersonating other existing channels and from posting content designed to appear as though it originates from someone else.[77] Reddit prohibits the impersonation of individuals or entities in a "misleading or deceptive manner."[78]

Given the nuances of each of these policies, whether they might reasonably be applied to the case at hand depends on several factors. Assuming the individual profiles that were used to contact Facebook page editors were fake accounts, Facebook's misrepresentation policy might apply. If they were operated by real individuals who sought not to promote Israeli content as claimed but to drive traffic to dubious websites for financial gain, Facebook's inauthentic behavior policy might still have been violated. The counterargument is that if the actors in question are Israeli and seek to promote pro-Israeli content, it is difficult to prove that their primary intent is to benefit financially and not truly to promote a viewpoint.

Twitter's policy on platform manipulation might apply insofar as the accounts engage in harmful behavior, but the sole act of creating fake or multiple accounts does not appear sufficient to violate the policy, as the platform recognizes the legitimacy of accounts created for expressing distinct identities, parodies, commentaries, or fan tributes.[79] Similarly, YouTube and Reddit make no presumption that its users are accurately representing their true personal identities on their channels, and individuals are free to operate multiple channels or accounts, as long as none of them impersonate other existing channels or accounts or falsely claim to be representative of some other actual person or organization.

Coordinating inauthentic behavior across platforms: Another possible opportunity for pursuing enforcement against the content mill's activities relates to the group's "coordinated inauthentic behavior," which can include the use of fake or misleading accounts. The specific term coordinated inauthentic behavior is used predominantly by Facebook, but Twitter, YouTube, and Reddit have similar policies in place.[80]

Facebook defines coordinated inauthentic behavior as the use of multiple accounts, groups, or pages to engage in inauthentic behavior. The key to this policy is that it involves not just a single instance of misrepresentation but multiple instances intended to work in tandem.[81] Considering that at least nineteen Facebook pages were found to be amplifying the content related to this influence operation, often seconds after a new post was added, it is likely that the group is engaging in such coordination and potentially in violation of this policy.[82] Facebook may be able to see if the same accounts are acting as administrators across multiple pages to coordinate posts, but that information is not publicly available.

Twitter's platform manipulation policy also prohibits inauthentic engagements and coordinated attempts to "artificially influence conversations through the use of multiple accounts, fake accounts, automation, and/or scripting."[83] Twitter's rules around automation, however, may complicate this argument, as Twitter does permit automated retweets and automated sharing of content. All the standard Twitter rules, including those related to spam and misrepresentation, apply when developing automation tools, but such tools are technically permissible.[84]

While Reddit's policies around the authenticity of content differ slightly from those of Facebook and Twitter, their policy on spam would seem to cover this type of behavior too. The platform prohibits repeatedly posting the same content across threads, a subreddit, or subreddits as well as the repeated posting of unrelated, off-topic, or link-farmed content.[85] YouTube's policy on repetitive content applies both to videos and to comments but similarly falls under its spam policy. The policies of these two platforms seem to place greater emphasis on the volume of content being published than the authenticity of the content.

The content mill's behavior does appear troubling under both policies that focus on authenticity and policies that focus on engagement volume. If the content is primarily distributed for financial gain under the guise of political concerns, then it lacks authenticity; meanwhile, if the content is distributed nearly instantly across multiple platforms, then it raises concerns about engagement volume.

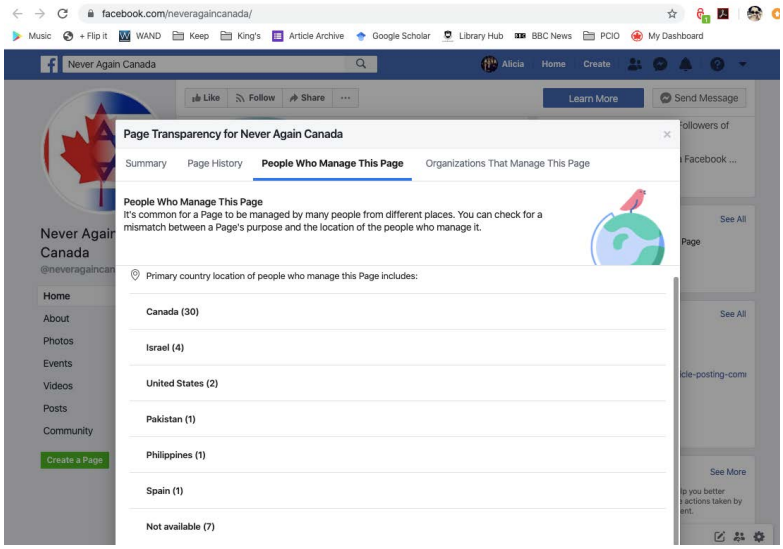
Taken together, these policies might address specific aspects of the content mill operators' problematic behavior. Facebook's enforcement priorities provide for the removal of content and the removal of accounts that repeatedly engage in problematic behavior, and its enforcement report on community standards claims that they prioritize fake account enforcement against those who "seek to cause harm and find many of these fake accounts are used in spam campaigns and are financially motivated."[86] Twitter also removes content and temporarily or permanently suspends accounts that engage in problematic behavior, while Reddit and YouTube ban users that repeatedly violate their respective terms of service.[87] All of the platforms ban individuals and accounts that have been suspended or terminated from creating new accounts to prevent them from seeking to circumvent bans.[88]

These policies provide platforms with several avenues for tackling the techniques used in the content mill's influence campaign. But the policies lack a holistic approach to assessing why certain content is problematic when combined with certain behaviors. The campaign in question is troubling not merely because actors are monetizing attention but because they are potentially misrepresenting themselves and disseminating content that relies on false or misleading information, polarizes political debates, and, therefore, pollutes the information environment.

For instance, in the case of the "Never Again Canada" Facebook group, the group's "page transparency" details (provided by Facebook) indicated that, while the vast majority of group administrators were located in Canada, there were sixteen others located outside the country (see screenshot 12). This raises questions about distinguishing between foreign and domestic actors in a hyperconnected world: who has the right to influence politics in a given geographic region?

This case study highlights the difficulty of applying one-dimensional policies to a multidimensional problem. Without such a coherent framework, an effective strategy for resolving this kind of campaign will be elusive. Too many fundamental questions remain about identifying authentic actors from inauthentic ones, distinguishing foreign users from domestic ones, and contextualizing actors' motives.

SCREENSHOT 12.



The domestic and foreign administrators of a Canadian Facebook group

The policies social media platforms have on the books are inadequate to fully deal with actors that use a variety of borderline-permissible techniques in coordination to achieve unfavorable outcomes. A better approach might be to consider how a single actor or group could use an array of problematic tactics to conduct influence operations whose effects violate the norms of platforms and, indeed, of democratic societies (at least). By analyzing the data related to individual policy enforcements more holistically over time, platforms may be able to identify patterns of problematic behavior that might be indicative of illegitimate influence operations. Such progress could enable online platforms to move against them before external researchers and media alert them of such issues.

Still, the case study highlights how platforms are asked to prioritize different levels of harm in their enforcement efforts. Although platform manipulation policies can and do cover the kinds of harms outlined above, they are also meant to deal with arguably more serious cases in which governments employ influence operations meant to shape an entire information space, not just drive clicks. At the same time, the persistence of the kind of harmful content distributed by the content mill operators could have a cumulatively negative effect by making it seem normal to encounter inflammatory,

misleading content that privileges a worldview associated with anti-Muslim sentiments. Online platforms have been put in a position to outline these enforcement priorities, absent adequate, coherent guidance from governments or civil society.

Treaties and International Law

There are few, if any, international laws or treaties governing influence operations, particularly when such activities are conducted by civilians or proxy organizations and not directly orchestrated by one state against another.[89] But there are important principles enshrined in international law that support some relevant national policies.

For example, the *UN Declaration on the Inadmissibility of Intervention and Interference in the Internal Affairs of States* asserts that "states and peoples" have the right "to have free access to information and to develop fully, without interference, their system of information and mass media."[90] This declaration goes on to mention that states will not conduct "any defamatory campaign, vilification or hostile propaganda for the purpose of intervening or interfering in the internal affairs of other states."[91] Moreover, it underscores that states have the right:

> to combat, within their constitutional prerogatives, the dissemination of false or distorted news which can be interpreted as interference in the internal affairs of other states or as being harmful to the promotion of peace, co-operation and friendly relations among states and nations.[92]

Likewise, Article 20 of the *UN International Covenant on Civil and Political Rights* prohibits "any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence,"[93] which much of the inflammatory content pushed by the influence operators in this case arguably attempts to do.

Nonetheless, given that the UN is a multilateral organization, it is unlikely that a sizable enough majority of its member states would judge the influence activities outlined in this case study to be in breach of either document. After all, there is no conclusive evidence that the operators behind this campaign are acting at the behest of the Israeli government. Moreover, targeted countries would need to deem these activities sufficiently dangerous and important to raise this subject at the highest level of international diplomacy. That being said, individual countries could pass new domestic legislation enshrining some of the language mentioned above to regulate online communications or attempt to rely on some existing national laws to do so in the ways outlined below.

National Laws

The content mill's online activities span several countries with differing legal systems, with actors in Israel targeting audiences in Australia, Canada, the UK, and the United States through what appear to be willing local partners. Examining the legal provisions of the various jurisdictions affected by this conduct may therefore provide insights into how the governments of these countries are seeking to deal with this kind of influence campaign, demonstrating perhaps that in some cases the principles underpinning existing laws can be translated in some respects to meet the threat at hand. Countries have proposed new legislation or used existing laws to deal with online harms, but many of these laws face implementation challenges because of the necessity of attribution and the problem of extraterritoriality. Because many of these laws attempt to blame actors or platforms for engaging in harmful behavior or failing to stop others from engaging in harmful behavior, these kinds of legal solutions may prove an ill-suited tool for countering influence operations. Comprehensive legislative approaches might instead encourage platforms to share information about online harms.

Australia

Australia has at least three relevant laws that might apply to the influence campaign in question. The first, the 2015 *Enhancing Online Safety Act*, created an e-safety commissioner position with the power to demand that social media companies take down harassing or abusive posts (as well as material that is otherwise illegal, such as content involving child abuse).[94] The powers under this act were expanded in 2018 to include revenge porn, a development with implications for deepfakes, which are "hyper-realistic video, audio, or images of someone appearing to do and say something they didn't do or say."[95] Yet this law tends to deal more with cyber bullying and targeted attacks than the type of general discriminatory speech exhibited in the case of the content mill, with the exception of the directed attack on Faruqi mentioned earlier.

The second law was adopted in 2019 in the wake of the Christchurch terrorist attacks. The *Sharing of Abhorrent Violent Material Act* lets the government fine social media companies and exposes executives to the risk of jail time if they fail to rapidly remove "abhorrent violent material" from their platforms.[96] While the content shared by the content mill operators in this case might be hateful and misleading, it likely does not meet the law's threshold for violent material.

Finally, Australia's 1975 *Racial Discrimination Act* has been used to address "offensive behaviour" based on race, color, or national or ethnic origin. The law prohibits acts conducted in a public place that are likely to "offend, insult, humiliate or intimidate another person or a group of people" based on the characteristics written above.[97] The law was applied in a 2011 case involving a political commentator and two articles that made offensive comments about members of the country's aboriginal

community, which the court said "were not written in good faith and contained factual errors."[98] The law defines a public place as "any place to which the public have access as of right or by invitation, whether express or implied and whether or not a charge is made for admission to the place."[99] Applying this law to the case of the content mill and others like it would require a determination that a social media platform constitutes a "public place" under this definition and that the content was harmful enough to be likely to cause offense, insult, humiliation, or intimidation. In one 2002 case, the Australian Federal Court adjudicated an appeal involving a website promoting anti-Semitic content and found that the site violated the *Racial Discrimination Act*, meaning that the law could apply to online content.[100]

Canada

Canada has one relevant law, the *Canadian Criminal Code*, that criminalizes some offenses related to hate speech and propaganda, but it does not define hate speech. Offenses related to hateful propaganda include "advocating genocide," "public incitement of hatred," and "willful promotion of hatred . . . against any identifiable group," including one distinguished by its religion.[101] The law stipulates that this public information must be published deliberately and not carelessly.[102] The content pushed through by the content mill's influence campaign arguably aimed to foster Islamophobic sentiments, actions that could be deemed the promotion of willful hatred against a religious group.

The problem is that, to prosecute a crime under this law, the perpetrator must be known. If they are neither located in Canada nor Canadian, such limitations likely would prevent the law from applying. A further hurdle in Canadian jurisprudence is that most of Canada's existing laws on false or misleading information (that is distinct from its legislation on hate speech) apply to radio or television broadcasters but exempt social media platforms and their users.[103]

Israel

Israel has at least three laws or bills that deal with propaganda, including messaging from foreign sources. These laws focus specifically on elections and would not apply to the circumstances of the content mill case.[104] Moreover, these laws are focused on the targeting of domestic Israeli audiences and do not necessarily govern propaganda activities undertaken by Israeli citizens targeting citizens in other countries. Israeli lawmakers proposed two other laws targeting social media companies in 2016 and 2018, bills that demanded the removal of any content that incites violence or acts of terrorism.[105] It is unclear if these legal provisions have been adopted. Given that the operators of this influence campaign have been attributed by other sources to Israelis targeting citizens in other countries,[106] it is unlikely that these laws, should they be passed, would prohibit this activity unless international pressure were applied to Israel to push its citizens to curtail this conduct.

United Kingdom

Though the UK has criminalized terrorist-related online content, its laws on hate crimes and discrimination are scattered throughout the legal code and many of them are too outdated to deal effectively with online conduct and social media platforms. The UK's 2019 *Counter-Terrorism and Border Security Act* made it illegal to post, share, or download terrorist-related material.[107] The content posted by the operators of the content mill in question, while troubling, falls outside the realm of terrorism. Sharing content from the Jewish Defense League would not qualify as sharing terrorist content in the UK, as the group is not on the government's list of "proscribed terrorist groups or organisations."[108]

That said, other British laws governing hate crimes and discrimination might apply.[109] Racial and religious hatred is defined under these laws "as hatred against a group of persons by reason of the group's colour, race, nationality (including citizenship) or ethnic or national origins or religion (or indeed perceived religion)." Should a person posting online be "threatening, abusive or insulting," it is possible that legal action might be taken against them.

However, even in the case of these laws, legal complications abound. This approach is highly subjective in terms of determining what speech might qualify, to say nothing of the need to know exactly who is behind the post and whether British law would even apply to them. Moreover, as one report from the Home Affairs Select Committee in the UK House of Commons noted, most of the relevant pieces of legislation that might govern hate speech and abuse are "spread across a number of different Acts of Parliament and each was passed before social media were mainstream tools, and some Acts were passed even before the internet itself was widely used."[110] These limitations make these laws woefully outdated for combating digital influence operations.

United States

U.S. constitutional provisions make it very difficult to take legal action against those who perpetuate hateful and inflammatory content, except in very narrow circumstances. The First Amendment to the U.S. Constitution renders prosecuting hate speech extremely difficult. This amendment protects free speech, including that which might be perceived as hateful, subject to very narrow exceptions.[111]

One group of exceptions involves speech that creates the realistic risk or threat of physical violence. For example, a true threat to physically harm someone can be prosecuted, as can incitement to *imminent* lawless action, or so-called fighting words—speech so offensive that it risks provoking an *immediate* breach of the peace (such as starting a fight). The thresholds for meeting each of these exceptions is very high and would almost certainly not apply to any of the content at issue here. After all, even general advocacy of violence is protected under the First Amendment.[112]

Other First Amendment exceptions might allow private parties to sue the operators of this content mill on various grounds. An individual falsely depicted in a social media post may be able to sue the operators for defamation, provided the author was negligent or reckless in posting the false information (which often appears to have been the case).[113] A legitimate news site whose language has been extensively copied verbatim—such as Radio Free Europe—could sue the operators for copyright infringement. (The aggrieved news site could also rely on the Digital Millennium Copyright Act to issue a takedown request to the platform or web host serving up the plagiarized content).[114]

In addition, a social media platform could perhaps sue the operators for breach of contract (due to terms of service violations) or perhaps fraud. Platforms are slowly beginning to experiment with civil suits against exceptionally bad actors. For example, Facebook recently sued a company that was cloaking the true destination of its links to trick Facebook users into installing malware.[115] However, like in the cases of other aforementioned national laws, any of these legal avenues would ultimately require identifying the operators. And if the operators do not reside in the United States or have U.S.-based assets, any legal judgment against them in U.S. courts would have limited effects.

Notably, U.S. law offers no mechanism for legally applying pressure to social media platforms for the removal of such content. The Communications Decency Act of 1996 might have been used to counter hate speech, but Section 230 seems to absolve social media companies of responsibility for content that other people publish on their platforms.[116]

In the case of influence operations involving foreign actors, the U.S. Foreign Agents Registration Act (FARA), which regulates the lobbying influence wielded by foreign actors over U.S. institutions, has recently been applied. FARA is a transparency statute that requires foreign agents to register with the U.S. government, report their activities, and conspicuously label informational materials as the work of a foreign agent.[117]

Though rarely used in recent decades, the law has experienced a revival in the last few years. Former special counsel Robert Mueller used FARA to charge the Russian operatives behind foreign influence activities discovered during the 2016 U.S. presidential election.[118] Given that there are clearly foreign operators engaging with U.S. audiences, FARA might be one avenue for prosecution if the operators continue refusing to register and label their material according to the statute's provisions. Yet the relatively novel use of FARA to combat social media influence activities faces the same hurdles as other legal avenues: it is challenging and resource-intensive to identify perpetrators and may be impossible to hold them accountable in U.S. courts.

Overall Observations

Many of these various national laws deal with two challenges: attribution and extraterritoriality. Using almost any of the legal tools outlined above would require attributing malign activity to a specific actor, a task that is often difficult to do when identities are easily obscured online. Furthermore, even when a reliable attribution can be made, it is often difficult to marshal the resources necessary to hold an actor accountable. Part of this problem stems from concerns about extraterritoriality—that is, whether a country's law applies to citizens outside of the country itself. In some cases, like in Australia, courts have ruled that when online content can be viewed by Australian citizens, it is subject to Australian law, regardless of where the creator resides or the data is hosted.[119] In practice, even this interpretation would likely be difficult to act on, as it may require extradition, and prosecutors may well decide to pursue less resource-intensive cases.

Of course, attempting to identify laws that deal specifically with influence operations presupposes that such laws do or should exist. On this point, expert opinions can differ, as it is too early to say whether reliance on existing laws is sufficient or whether new, more comprehensive laws are needed. One way of assessing some of the laws that could apply to influence operations is to say that policymakers are ill-equipped to use legal instruments to this end, given the slow pace of legal change, the difficulty of attribution, and the uncertainty surrounding which laws might, in practice, be most appropriate for countering influence operations. Because many of these laws focus on criminalization, rather than promoting information sharing or transparency, they fall short of the goal of creating a healthy online environment, even when they do apply. However, another way of looking at things might be to recognize that a comprehensive legislative solution that attempts to delineate what is acceptable or unacceptable content is ill-suited to dealing with influence operations. Before putting forward new, comprehensive laws to deal with influence operations, policymakers may favor laws that emphasize transparency and disclosure, rather than criminalization, to help better understand and manage the proper balance between state-driven and platform-driven solutions.[120]

Those caveats notwithstanding, the creative ways that policymakers have been applying existing law to influence operations—like the use of FARA in the United States—may suggest that the principles related to inauthenticity and harmful content that underpin influence operations can be found in existing law and applied when sufficient harm has been done and resources can be mustered.[121]

Conclusion

The case of this content mill highlights the many challenges facing researchers, social media platforms, and governments trying to understand how best to counter influence operations.

First and foremost is the question of what precisely, if anything, is most troubling about what online actors like the operators of this content mill are doing. Pertinent fundamental questions include whether the behavior demonstrated and the content promoted by the operators are truly problematic, if so why that behavior and content are worrisome, and who gets to make those determinations. There is an argument to be made that the operators are simply using, rather than abusing, online social network platforms. The fact that divisive content can be monetized may not be a flaw as much as a feature of the design of social media platforms. As of now, at least one of the authors of this study remains deeply ambivalent about whether or not the described behavior should be considered problematic, and it would not be a surprise to hear that many in the influence operations community are wrestling with similar thoughts. At the very least, there is not yet a consensus on what constitutes appropriate online behavior in such situations.

Much of the media coverage of this influence operation was problematic. News coverage did not adequately clarify that national statutes and international law do not regulate many or most of the objectionable activities in question. In the relative absence of suitable government regulatory frameworks, media organizations tend to simply assert or imply that online platforms must contain or stop such behavior. But media coverage of this case and others like it all too often has not explored the sorts of enforcement dilemmas that platforms face. News consumers are, as a result, being unrealistically primed to expect platforms simply to identify and stop the behavior.

Media pressure, in tandem with unclear guidance from governments and civil society, forces social media platforms to draw lines in the sand and to grapple with difficult questions: Is it acceptable for foreign actors to work in tandem with domestic administrators of online communities in other countries? Is omission a form of disinformation when it comes to covering such news stories responsibly? Is it acceptable to make money pushing politically charged content? What characteristics push some content over the line into hate speech? Many of these questions were contested even in an analog environment, and they have only become thornier in the age of digital media.

More specifically, this case study exposes a few limits in the policies of online platforms related to inherent ambiguities in online discourse, including the difficulty of determining motives, especially when dealing with anonymous actors. The way the operators have been using fake accounts to gain access to editor privileges on Facebook pages raises questions about distinguishing between foreign

and domestic actors and assessing the intent behind actors' online behavior. As others have pointed out, the fact that the owners of these pages and groups are citizens of the countries against which the content is being directed complicates the platforms' efforts to determine authenticity, which is at the heart of many of their policies related to influence operations.[122] Their decision to cede control of these pages highlights the risk that operators of this and other content mills could co-opt legitimate actors as proxies to promote their agendas.

Even before posing questions about the policies of social media platforms, there are more fundamental matters at stake: what is it about efforts to shape public perceptions on issues of international concern that makes certain behavior problematic? If it is deceptive practices, then the focus returns to the difficulty of assessing the intent of online actors. How can corporate actors and government officials be sure that the primary motivation of a given network is financial gain and not influence over international public opinion? How important is it to distinguish these motives if the real-world impact of the behavior is the same regardless of why the content is promoted?

Assuming the *behavior* is problematic, social media platforms must gauge whether such conduct is appropriate largely by measuring authenticity or engagement volume. But absent any real way of determining the operators' motives, authenticity is an inherently unidentifiable characteristic. The content of the posts and the behavior of the operators point to mixed motives—the operators may truly believe in the content they are sharing and may simply be fortunate to have found a way to monetize their deeply held beliefs; on the other hand, the operators may also be exploiting divisive political issues primarily for financial gain. In many cases, it can be difficult to tell.

Assuming the *content* is problematic, social media companies often look for explicit calls for violence, threats, or otherwise harmful content before they will take action. These platforms have already devised technical and human solutions for dealing with actors that commit clear offenses.[123] But sophisticated, malign online actors are aware of platform standards and generally avoid committing clear offenses to ensure that their content is not taken down, employing a constantly shifting repertoire of techniques.

Social media platforms' piecemeal approaches to dealing with individual aspects of influence operations raise questions about whether these platforms are looking at this problem systemically and whether government officials and civil society are giving them the guidance they need to prioritize certain harms over others. Surely, each type of policy infringement by malign influence operations generates a wealth of data over time that platforms can use to identify and pinpoint behavioral patterns associated with such campaigns. Are social media companies, given their preference for focusing on individual policies, holistically connecting all the dots?

The answer appears to be no. As this case study shows, social media platforms, at least as far as their public-facing policies are concerned, are focused more on individual activities and not the sum of objectionable behaviors that comprise influence operations. Assessing the actors, motives, and techniques in isolation precludes these platforms from robustly analyzing the interconnections that could make influence operations easier to detect.

A more holistic approach would entail, first and foremost, developing metrics to understand the overall impact of influence operations; exploring potentially effective countermeasures; and sharing that information with researchers, civil society members, and government officials. One of the largest hurdles to effectively countering influence operations is the lack of an agreed-upon methodology for measuring and evaluating impact. Better data would help platforms make sense of contextual factors, repeated behaviors, and severity of harm.

Beyond the responses of online platforms themselves, some national laws that might be applicable are constrained by a traditional framework that requires clear acts of deliberate criminal activity to be reliably attributed to their perpetrators. This very narrow framing leaves a lot to be desired, especially given how easily actors can obfuscate their true identities and locations. Traditional criminal legal systems may not be well equipped to deal with influence operations. Nor is it clear that they should be.

There are reasons why liberal democracies, which are particularly susceptible to influence operations and highly value principles of free expression, have so far refrained from enacting more stringent legislation to combat them. The consequences of potential missteps are too high a price to pay in the eyes of many citizens. And legislative approaches in other parts of the world have backfired and in some cases allowed governments in some countries (including Singapore) to take action against opposition figures under the guise of such laws.[124] Alternative approaches, such as norms or laws that deal with disclosure and transparency, might be better suited in such situations, but those standards would also need to be articulated.

While online platforms have borne the brunt of much of the public pressure to deal with influence campaigns, many of these questions are also important for society in general to contemplate and should not be answered by industry actors alone. In the rush to do something about influence operations, societies might do well to first convene a coalition of cross-sector, multidisciplinary stakeholders to determine what the lines in the sand are. Some of this work is already under way: operational researchers, academics, think tank scholars, and others have devoted considerable time and effort to studying influence operations and starting to address relevant policy questions.[125] These

stakeholders might also take pains to better understand the impact of online content on those who engage with it. While metrics about how many users engage with a given piece of content are available, they say little about whether or how such engagement affects individuals' beliefs or behavior. Any approach to dealing with problematic behavior or problematic content should be based on an understanding of the potential consequences of its spread.

Once stakeholders have achieved a foundational understanding of the impact of such influence operations, have developed greater consensus, and have established guiding principles or practices, private- and public-sector entities should test their responses in adversarial exercises. Because the tactics of influence operators are constantly shifting, any overly rigid attempt to draw lines in the sand will be met immediately with threat actors that want to wipe those lines away. The danger in publicly promulgating a policy is that doing so gives threat actors a clear sense of what techniques must be avoided to remain undetected. Rather than relying on the presumption that users will interpret standards in the broadest possible sense, platforms should assume that users will interpret them in the narrowest sense and that fair enforcement will do the same.

This case study admittedly highlights more questions than it answers. Nonetheless, it is intended to reveal the difficult quandaries that citizens, social media platforms, and governments face as influence campaigns inspired by often unclear motives continue to prey on social divisions. Answering such questions will require the concerted efforts of this full range of stakeholders.

About the Authors

Elise Thomas is a freelance journalist and researcher working with the International Cyber Policy Center at the Australian Strategic Policy Institute. Her writing has been published in *Wired*, *Foreign Policy,* the *Daily Beast*, the *Guardian Australia*, SBS, *Crikey*, and the *Interpreter* of the Lowy Institute.

Natalie Thompson is a James C. Gaither Junior Fellow with the Technology and International Affairs Program at the Carnegie Endowment for International Peace.

Alicia Wanless is the co-director of the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace and a doctoral researcher in war studies at King's College London.

Notes

1 Samanth Subramanian, "Inside the Macedonian Fake-News Complex," *Wired,* February 15, 2017, https://www.wired.com/2017/02/veles-macedonia-fake-news/.
2 "Never Again Canada," Facebook, archived February 19, 2020, http://archive.is/vTlWc.
3 Craig Silverman, Jane Lytvynenko, Marco Chown Oved, and Alex Boutilier, "How a Facebook Page Dedicated to Fighting Anti-Semitism Became a Hub for Anti-Muslim Content," Buzzfeed News, April 12, 2019, https://www.buzzfeednews.com/article/craigsilverman/never-again-canada-jdl-facebook.
4 Facebook, "Never Again Canada."
5 Silverman, et al., "How a Facebook Page Dedicated to Fighting Anti-Semitism Became a Hub for Anti-Muslim Content."
6 Ibid.
7 Ibid.
8 Christopher Knaus and Michael McGowan, "Far-Right 'Hate Factory' Still Active on Facebook Despite Pledge to Stop It," *Guardian*, February 4, 2020, https://www.theguardian.com/australia-news/2020/feb/05/far-right-hate-factory-still-active-on-facebook-despite-pledge-to-stop-it.
9 "Arab-Muslim Father Terrorizes His Infant Son With Gunshots," Freespeechfront.com, September 2019, http://www.free-speechfront.info/2019/09/arab-muslim-father-terrorizes-his.html, archived on February 13, 2020, http://archive.is/T3H1t.
10 Ramadan Al Sherbini, "Outrage After Saudi Man Puts Gun in Baby's Mouth," *Gulf News,* September 15, 2019, https://gulfnews.com/world/gulf/saudi/outrage-after-saudi-man-puts-gun-in-babys-mouth-1.66437783.
11 These findings were made using the proprietary tool Buzzsumo to analyze the relevant domains on February 17, 2020. For more information on the tool, see the Buzzsumo website. "Buzzsumo," Buzzsumo, https://buzzsumo.com/
12 "Shock as Bulgaria, Romania, Serbia and Greece Declare War Against Radical Islam, the UN and the EU as They Join Forces With Israel," Freespeechfront.net, November 2018, http://www.freespeechfront.net/2018/11/shock-as-bulgaria-romania-serbia-and.html, archived on February 17, 2020, http://archive.is/p4jlN.
13 "PM Netanyahu's Statement at the Craiova Forum Meeting," Israeli Prime Minister's Office, November 2, 2018, https://www.gov.il/en/departments/news/event_forum021118; and Vuk Vuksanovic, "Israel Discovers Europe's Soft Underbelly—the Balkans," Middle East Eye, December 5, 2018, https://www.middleeasteye.net/opinion/israel-discovers-europes-soft-underbelly-balkans.
14 "PM Netanyahu's Speech to the Christian Media Summit at the Israel Museum in Jerusalem," Israeli Prime Minister's Office, October 15, 2017, https://www.gov.il/en/departments/news/speechmuseum151017.
15 "Watch: Angela Merkel Complains 'Trump Is Destroying the UN,'" Politicsonline.net, January 1, 2019, http://www.politicsonline.net/2019/03/watch-angela-merkel-complains-trump-is.html, archived on February 17, 2020, http://archive.is/1mewT.
16 The statistics were collected from Buzzsumo. Buzzsumo, February 26, 2020, https://app.buzzsumo.com/content/web?begin_date=Feb%2017%202019&end_date=Feb%2017%202020&q=http%3A%2F%2Fwww.freespeechfront.net%2F2018%2F11%2Fshock-as-bulgaria-romania-serbia-and.html&result_type=total; Tony Stirrat, @anthonystirrat, Twitter post, October 11, 2019, 4:02 a.m., https://twitter.com/anthonystirrat/status/1182567149627498497; and Richard James, @Skisidjames, Twitter post, December 7, 2018, 2:27 p.m., https://twitter.com/skisidjames/status/1071124145935712257.

17 "Shock as Bulgaria, Romania, Serbia and Greece Declare War Against Radical Islam, the UN and the EU as They Join Forces With Israel," Beforeitsnews.com, December 10, 2018, https://beforeitsnews.com/alternative/2018/12/shock-as-bulgaria-romania-serbia-and-greece-declare-war-against-radical-islam-the-un-and-the-eu-as-they-join-forces-with-israel-3651418.html; and "Shock as Bulgaria, Romania, Serbia and Greece Declare War Against Radical Islam, the UN and the EU as They Join Forces With Israel," Nesaranews.blogspot.com, December 10, 2018, http://nessaranews.blogspot.com/2018/12/shock-as-bulgaria-romania-serbia-and.html.

18 Social Insider, https://app.socialinsider.io/.

19 Guardians of Australia, Facebook post, February 14, 2020, 9:00 a.m., https://www.facebook.com/GUARDAUSTRALIA/posts/2535884956651072; and Poiticsonline.net, "Girl, 6, Is Forced to Marry a 55-Year-Old Imam in Exchange for a Goat in Afghanistan," October 2019, http://www.thepolitics.online/2019/10/girl-6-is-forced-to-marry-55-year-old.html?fbclid=IwAR0SDqHLFEstyy6Vnl0O1x5rSrCh0rz9b4Gjb5OXAbUxfl3O3j09ZG_t89U, archived on February 26, 2020, https://archive.is/1tYwS.

20 "Afghan Father Sells 6-yo Daughter Into Marriage for a Goat, Bag of Rice," RT, August 6, 2016, https://www.rt.com/news/354832-afghanistan-girl-sold-marriage/; and David Icke, "Afghan Father Sells Six Year-Old Daughter Into Marriage for a Goat and a Bag of Rice," Davidicke.com, August 7, 2016, https://www.davidicke.com/article/381073/afghan-father-sells-six-year-old-daughter-marriage-goat-bag-rice.

21 Pamela Constable and Sayed Salahuddin, "Six-Year-Old Afghan Girl Reportedly Sold in Marriage," *Washington Post*, July 31, 2016, https://www.washingtonpost.com/world/asia_pacific/six-year-old-afghan-girl-reportedly-sold-in-marriage/2016/07/31/a71e9a30-574e-11e6-9767-f6c947fd0cb8_story.html.

22 Never Again Canada, Facebook post, archived October 27, 2019, https://web.archive.org/web/20191027212302/https://www.facebook.com/neveragaincanada/.

23 "Watch: Adult Iranian Muslim Man Marries a 11-Year-Old Girl," Thepolitics.online, September 2019, http://www.thepolitics.online/2019/09/watch-adult-iranian-muslim-man-marries.html?fbclid=IwAR2gWsEwmF9Rwrs9AVn1G1gHwmT93fgWJCBltmrGdpY7FP-v0DpHbHgqJRs, archived on February 26, 2020, https://archive.is/zHM9W.

24 "Watch: Adult Iranian Muslim Man Marries a 11-Year-Old Girl," Trumptrain.com, October 13, 2019, https://www.trump-train.com/2019/10/watch-adult-iranian-muslim-man-marries.html, archived on February 26, 2020, https://archive.is/U0efN.

25 Golnaz Esfandiari, "Child Bride: 11-Year-Old Iranian Girl's Marriage Annulled After Public Outcry," Radio Free Europe/Radio Liberty, September 4, 2019, https://www.rferl.org/a/iran-child-bride-marriage-annulled-outcry-11-year-old/30146652.html.

26 The information in figure 1 was compiled by the authors based on their analysis of the web domains in question. See Google Analytics Help, "Tracking ID and Property Number," https://support.google.com/analytics/answer/7372977; and Google Analytics Help, "Property," https://support.google.com/analytics/answer/6086081?hl=en&ref_topic=6083659.

27 "Cookie and Privacy Policy," Speechpoint.net, http://www.speech-point.net/p/cookie-and-privacy-policy.html, archived on February 27, 2020, http://archive.is/RYFt2.

28 "Cookie and Privacy Policy," Speechpoint.net; "Cookie and Privacy Policy," Speechline.net, http://www.speechline.net/p/cookie-and-privacy-policy.html, archived on April 22, 2020, http://archive.is/eyDZu; "Cookie and Privacy Policy," Freespeechfront.com, http://www.free-speechfront.info/p/cookie-and-privacy-policy.html, archived on April 22, 2020, http://archive.is/zpSdm; "Cookie and Privacy Policy," Thepolitics.online, http://www.thepolitics.online/p/cookie-and-privacy-policy.html, archived on April 22, 2020, http://archive.is/DhNbM; "Cookie and Privacy Policy," Speechpoint.com, http://www.speech-point.com/p/cookie-and-privacy-policy.html, archived on April 22, 2020, http://archive.is/uk2Xw; "Cookie and Privacy Policy," Freespeechfront.com, http://www.freespeechfront.net/p/cookie-and-privacy-policy.html, archived on April 22, 2020, http://archive.is/jiYHS.

29 "U.S. Cuts All Funding and Announces Withdrawal From U.N.'s Cultural Agency Over Pro-Islamic Bias," Politicsonline.net, April 4, 2019, archived on February 10, 2020, http://archive.is/wip/Ld2lA.

30 No Sharia Law - Never Ever Give Up Australia, Facebook post, November 21, 2018, archived on February 10, 2020, http://archive.is/dfpQG.

31 "About Us," Thepolitics.online, http://www.thepolitics.online/p/about-us.html, archived on February 10, 2020, http://archive.is/9qGgB.

32 "Uberlink," Uberlink, https://uberlink.com/.

33 Aftab Ahmed, "Indian Police Detain Hundreds After Hindu-Muslim Clashes in New Delhi," *Globe and Mail,* February 28, 2020, https://www.theglobeandmail.com/world/article-indian-police-arrests-over-500-for-delhi-sectarian-violence/.

34 These calculations are based on the numbers presented in figure 1.

35 Richard Marriott, "49 Free Web Directories for Building Backlinks," Clambr.com, http://www.clambr.com/49-free-web-directories-for-building-backlinks/.

36 "Free Speech Time," Facebook post, archived on November 7, 2018, https://web.archive.org/web/20181107224637/https:/www.facebook.com/FreeSpeechTime/; "We Love Israel," Facebook post, archived on April 10, 2018, https://web.archive.org/web/20180410000014/https:/www.facebook.com/WeLoveIsraelPage/.

37 Knaus and McGowan, "Inside the Hate Factory."

38 Ibid.

39 Ibid.

40 Silverman, et al., "How A Facebook Page Dedicated To Fighting Anti-Semitism Became A Hub For Anti-Muslim Content."

41 Knaus and McGowan, "Inside the Hate Factory."

42 Ibid.

43 Ibid.

44 Facebook, "Guardians of Australia," Facebook Community, https://www.facebook.com/GUARDAUSTRALIA/.

45 There is a notable symmetry between the "Alisa" and "lizalol" Reddit accounts and the suspicious "Alicia" and "Liza" Twitter accounts. The lizalol151665 account has shared at least two articles from other sources, namely the *Jerusalem Post* and the *Daily Caller*, a right-wing site.

46 u/andynushil, Reddit account, archived on April 28, 2020, https://archive.is/GjOQx.

47 u/andynushil, Reddit post, archived on February 12, 2020, http://archive.is/Yy5cg.

48 "Free Speech Watch," YouTube, https://www.youtube.com/channel/UCIUIi4kCf7qNZZrdS5FQ0nQ/videos, archived on February 12, 2020, http://archive.is/DptrT.

49 "About," Gab, https://gab.com/about.

50 See Andy Nush, Gab, archived February 5, 2020, http://archive.is/Dsxu5. This account's first two posts, on May 20, 2019, shared posts from the right-wing conspiracy-theory-promoting account PrisonPlanet.

51 Rachel Rose, Gab, archived on February 5, 2020, http://archive.ph/qk9d3.
52 Website Terms of Service," Gab, https://gab.com/about/tos.
53 Ibid.
54 Jane Coaston, "Gab, the Social Media Platform Favored by the Alleged Pittsburgh Shooter, Explained," *Vox*, October 29, 2018, https://www.vox.com/policy-and-politics/2018/10/29/18033006/gab-social-media-anti-semitism-neo-nazis-twitter-facebook.
55 Silverman, et al., "How a Facebook Page Dedicated to Fighting Anti-Semitism Became a Hub For Anti-Muslim Content."
56 Twitter, "Terrorism and Violent Extremism Policy," Twitter Help Center, March 2019, https://help.twitter.com/en/rules-and-policies/violent-groups; and YouTube, "Violent Criminal Organizations," YouTube Help, https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436.
57 YouTube, "Hate Speech Policy," YouTube Help, https://support.google.com/youtube/answer/2801939?hl=en; Facebook, "12. Hate Speech," Facebook Community Standards, https://www.facebook.com/communitystandards/hate_speech; and Twitter, "Hateful Conduct Policy," Twitter Help Center, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.
58 Facebook, "9. Bullying and Harassment," Facebook Community Standards, https://www.facebook.com/communitystandards/bullying.
59 Facebook, "1. Violence and Incitement," Facebook Community Standards, https://www.facebook.com/communitystandards/credible_violence; and Twitter, "Violent Threats Policy," Twitter Help Center, March 2019, https://help.twitter.com/en/rules-and-policies/violent-threats-glorification.
60 Facebook, "12. Hate Speech."
61 Ibid.
62 Ibid.
63 Twitter, "Hateful Conduct Policy."
64 Twitter, "Abusive Behavior," Twitter Help Center, https://help.twitter.com/en/rules-and-policies/abusive-behavior.
65 YouTube, "Hate Speech Policy."
66 Reddit, "Do Not Threaten, Harass, or Bully," Account and Community Restrictions, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-threaten-harass-or-bully.
67 Knaus and McGowan, "Inside the Hate Factory."
68 Facebook, "1. Violence and Incitement"; Facebook, "12. Hate Speech"; Facebook, "9. Bullying and Harassment"; Twitter, "Violent Threats Policy"; Twitter, "Abusive Behavior"; Reddit, "Do Not Threaten, Harass, or Bully"; YouTube, "Hate Speech Policy"; YouTube, "Harassment and Cyberbullying Policy," YouTube Help, https://support.google.com/youtube/answer/2802268?hl=en&ref_topic=9282436.
69 Knaus and McGowan, "Inside the Hate Factory."
70 Twitter, "Our Range of Enforcement Options," Twitter Help Center, https://help.twitter.com/en/rules-and-policies/enforcement-options.
71 Facebook, "18. Spam," Facebook Community Standards, https://www.facebook.com/communitystandards/spam.
72 Reddit, "What Constitutes Spam? Am I a Spammer?" Account and Community Restrictions, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/what-constitutes-spam-am-i-spammer.
73 Twitter, "Platform Manipulation and Spam Policy," Twitter Help Center, September 2019, https://help.twitter.com/en/rules-and-policies/platform-manipulation.

74 Facebook, "17. Misrepresentation," Facebook Community Standards, https://www.facebook.com/communitystandards/misrepresentation.
75 Facebook, "20. Inauthentic Behavior," Facebook Community Standards, https://www.facebook.com/communitystandards/inauthentic_behavior.
76 Twitter, "Platform Manipulation and Spam Policy."
77 YouTube, "Policy on Impersonation," YouTube Help, https://support.google.com/youtube/answer/2801947?hl=en&ref_topic=9282365.
78 Reddit, "Do Not Impersonate an Individual or Entity," Account and Community Restrictions, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-impersonate-individual-or.
79 Twitter, "Platform Manipulation and Spam Policy."
80 Knaus and McGowan, "Inside the Hate Factory."
81 Facebook, "20. Inauthentic Behavior."
82 Silverman, et al., "How a Facebook Page Dedicated to Fighting Anti-Semitism Became a Hub for Anti-Muslim Content."
83 Twitter, "Platform Manipulation and Spam Policy."
84 Twitter, "Automation Rules," Twitter Help Center, updated November 3, 2017, https://help.twitter.com/en/rules-and-policies/twitter-automation.
85 Reddit, "What Constitutes Spam? Am I a Spammer?"
86 Facebook, "Community Standards Enforcement Report," Facebook Transparency, https://transparency.facebook.com/community-standards-enforcement; and Facebook, "Community Standards Enforcement Report: Fake Accounts," Facebook Transparency, https://transparency.facebook.com/community-standards-enforcement#fake-accounts.
87 Twitter, "Our Range of Enforcement Options," Twitter Help Center, https://help.twitter.com/en/rules-and-policies/enforcement-options; Reddit, "Reddit Content Policy," https://www.redditinc.com/policies/content-policy; and YouTube, "Community Guidelines Strike Basics," YouTube Help, https://support.google.com/youtube/answer/2802032.
88 Facebook, "17. Misrepresentation; Twitter, "Our Approach to Policy Development and Enforcement Philosophy," Twitter Help Center, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy; YouTube, "Channel Terminations," YouTube Help, https://support.google.com/youtube/answer/2802168?hl=en&ref_topic=9387060; and Reddit, "What is Ban Evasion?" Account and Community Restrictions, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/what-ban-evasion.
89 Trishana Ramluckan, Brett van Niekerk, and Alicia Wanless, "Cyber-Influence Operations: A Legal Perspective" (paper presented at the eighteenth European Conference on Cyber Warfare and Security, Coimbra, Portugal, July 2019), https://search.proquest.com/docview/2261019847/fulltext/339E344430F44477PQ/1?accountid=40995.
90 United Nations, "Declaration on the Inadmissibility of Intervention and Interference in the Internal Affairs of States," United Nations General Assembly A/RES/36/103, January 20, 1982, https://digitallibrary.un.org/record/27066?ln=en, 79.
91 Ibid., 80.
92 Ibid.
93 UN General Assembly, "International Covenant on Civil and Political Rights," December 16, 1966, United Nations Treaty Series Online, registration no. I-14668, https://treaties.un.org/doc/publication/unts/volume%20999/volume-999-i-14668-english.pdf, 171.

94 Australian eSafety Commissioner, "Our Legislative Functions," https://www.esafety.gov.au/about-us/who-we-are/our-legislative-functions.
95 "Deepfakes," Carnegie Endowment for International Peace, https://carnegieendowment.org/siliconvalley/deepfakes; Australian Department of Infrastructure, Transport, Regional Development and Communications, "Stronger Laws to Stop Image-Based Abuse," September 6, 2018, https://www.communications.gov.au/departmental-news/stronger-laws-stop-image-based-abuse; and Australian eSafety Commissioner, "Adult Cyber Abuse," https://www.esafety.gov.au/key-issues/adult-cyber-abuse.
96 Australian Attorney-General's Department, "Abhorrent Violent Material," https://www.ag.gov.au/Crime/Pages/abhorrent-violent-material.aspx.
97 Australian Federal Register of Legislation, *Racial Discrimination Act,* Part IIA.18C, 1975, https://www.legislation.gov.au/Details/C2016C00089.
98 "Bolt Breached Discrimination Act, Judge Rules," ABC News, September 28, 2011, https://www.abc.net.au/news/2011-09-28/bolt-found-guilty-of-breaching-discrimination-act/3025918.
99 Australian Federal Register of Legislation, *Racial Discrimination Act,* Part IIA.18C.
100 Australian Human Rights Commission, "Racial Vilification Law in Australia," October 2002, https://www.humanrights.gov.au/our-work/racial-vilification-law-australia#relevant. See the section on the 2002 case *Jones v. Toben*.
101 Canadian Government, *Canadian Criminal Code,* Part VIII Offences Against the Person and Reputation, Defamatory Libel, (R.S.C., 1985, c. C-46), Justice Laws Website, https://laws-lois.justice.gc.ca/eng/acts/C-46/page-68.html#docCont.
102 David Butt, "Canada's Law on Hate Speech Is the Embodiment of Compromise," *Globe and Mail*, May 12, 2018, https://www.theglobeandmail.com/opinion/canadas-law-on-hate-speech-is-the-embodiment-of-compromise/article22520419/.
103 Library of Congress, "Initiatives to Counter Fake News: Canada," last updated June 11, 2019, https://www.loc.gov/law/help/fake-news/canada.php.
104 Library of Congress, "Initiative to Counter Fake News: Israel," late updated June 11, 2019, https://www.loc.gov/law/help/fake-news/israel.php. Specifically, see the sections on the Committee for Examination of the Elections (Modes of Propaganda) Law (2015); the Elections (Modes of Propaganda) (Amendment No. 34) Bill, 5778-2018, (2018); and the Bill Addressing Foreign Propaganda (2018).
105 Gwen Ackerman, "Israel Set to Approve 'Facebook Law' Against Web Incitement," *Bloomberg*, July 16, 2018, https://www.bloomberg.com/news/articles/2018-07-16/israel-set-to-approve-facebook-law-against-web-incitement.
106 Knaus and McGowan, "Inside the Hate Factory."
107 UK Parliament, *Counter-Terrorism and Border Security Act 2019,* February 12, 2019, http://www.legislation.gov.uk/ukpga/2019/3/2019-04-12/data.xht?view=snippet&wrap=true.
108 UK Home Office, "Proscribed Terrorist Groups or Organisations," last updated November 4, 2019, https://www.gov.uk/government/publications/proscribed-terror-groups-or-organisations--2.
109 UK Crown Prosecution Service, "Hate Crime," https://www.cps.gov.uk/hate-crime; and UK Government "Discrimination: Your Rights," https://www.gov.uk/discrimination-your-rights.
110 UK House of Commons Home Affairs Select Committee, "Hate Crime: Abuse, Hate and Extremism Online," April 27, 2017, https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60904.htm.
111 Christopher Woolf, "Why America Doesn't Ban Hate Speech," *Public Radio International*, November 22, 2016, https://www.pri.org/stories/2016-11-22/why-america-doesn-t-ban-hate-speech.

112 Kathleen Ann Ruane, "Freedom of Speech and Press: Exceptions to the First Amendment," Congressional Research Service, September 8, 2014, 3–5, https://fas.org/sgp/crs/misc/95-815.pdf.
113 Ibid., 21.
114 Digital Media Law Project, "Protecting Yourself Against Copyright Claims Based on User Content," https://www.dmlp.org/legal-guide/protecting-yourself-against-copyright-claims-based-user-content.
115 Jessica Romero and Rob Leathern, "Taking Action Against Ad Fraud," Facebook Newsroom, December 5, 2019, https://about.fb.com/news/2019/12/taking-action-against-ad-fraud/.
116 U.S. Government, *Communications Decency Act, U.S. Code 47* (1996), Section 230.
117 U.S. Department of Justice, "General FARA Frequently Asked Questions," updated August 21, 2017, https://www.justice.gov/nsd-fara/general-fara-frequently-asked-questions.
118 Joshua R. Fattal, "The Justice Department's New, Unprecedented Use of the Foreign Agents Registration Act," *Lawfare*, December 18, 2019, https://www.lawfareblog.com/justice-departments-new-unprecedented-use-foreign-agents-registration-act.
119 Australian Human Rights Commission, "Racial Vilification Law in Australia." See the section on *Dow Jones Company Inc. v. Gutnick*.
120 For one proposal on such a legal framework, see Mark MacCarthy, "Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry," University of Amsterdam Institute for Information Law Transatlantic Working Group on Content Moderation Online and Freedom of Expression," February 12, 2020, https://www.ivir.nl/publicaties/download/Transparency_MacCarthy_Feb_2020.pdf.
121 Thank you to David O'Brien at the Berkman Klein Center for Internet and Society for sharing this important point. Conversation with Alicia Wanless, April 5, 2020.
122 Knaus and McGowan, "Inside the Hate Factory."
123 Facebook, "Community Standards Enforcement Report: Hate Speech," Facebook Transparency, https://transparency.facebook.com/community-standards-enforcement#hate-speech.
124 James Griffiths, "Singapore Just Used Its Fake News Law. Critics Say It's Just What They Feared," CNN, November 30, 2019, https://www.cnn.com/2019/11/29/media/singapore-fake-news-facebook-intl-hnk/index.html.
125 Researchers at Graphika, for example, regularly publish reports on influence operations based on the company's analysis of online content. For more, see Graphika, "Graphika Reports," https://graphika.com/reports. As for academics, just one example is the Misinformation Review run out of Harvard's Kennedy School of Government, an effort that was recently established to provide timely, peer-reviewed, interdisciplinary research on misinformation. See Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy, "The Harvard Kennedy School Misinformation Review," https://shorensteincenter.org/about-us/areas-of-focus/misinformation/the-misinformation-review/. In terms of think tanks, the Carnegie Endowment for International Peace's Partnership for Countering Influence Operations was established for this explicit purpose. Others that are researching and analyzing these issues include the Atlantic Council's Digital Forensics Research Lab, the German Marshall Fund's Alliance for Securing Democracy, and the RAND Corporation's Truth Decay Initiative, among many others. See, for instance, Atlantic Council, "Digital Forensics Research Lab," https://www.atlanticcouncil.org/programs/digital-forensic-research-lab/; German Marshall Fund, "Alliance for Securing Democracy," https://securingdemocracy.gmfus.org/; and RAND Corporation, "Countering Truth Decay," https://www.rand.org/research/projects/truth-decay.html.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

1779 Massachusetts Avenue NW | Washington, DC 20036 | P: + 1 202 483 7600

CarnegieEndowment.org