May 2021

Threat Report
The State of Influence Operations 2017-2020

FACEBOOK

TABLE OF CONTENTS

Executive Summary

Over the past four years, industry, government and civil society have worked to build our collective response to influence operations ("IO"), which we define as "coordinated efforts to manipulate or corrupt public debate for a strategic goal."

The security teams at Facebook have developed policies, automated detection tools, and enforcement frameworks to tackle deceptive actors — both foreign and domestic. Working with our industry peers, we've made progress against IO by making it less effective and by disrupting more campaigns early, before they could build an audience. These efforts have pressed threat actors to shift their tactics. They have — often without success — moved away from the major platforms and increased their operational security to stay under the radar.

Historically, influence operations have manifested in different forms: from covert campaigns that rely on fake identities to overt, state-controlled media efforts that use authentic and influential voices to promote messages that may or may not be false. In tackling these different problems, we rely on distinct policies and enforcement measures.

For example, we clearly label state media so people can know who's behind the content they see and judge its trustworthiness. We limit less sophisticated, inauthentic efforts to cheat our systems by applying scaled enforcement, including: warnings, down-rankings, and removals.[1] If we find that *content* itself violates our Community Standards (*e.g. harmful health-related misinformation or voter suppression misinformation*, we remove it) or reduce its distribution.

However, when a threat actor conceals their identity through deceptive *behavior*, the public will lack sufficient signals to judge who they are, how trustworthy their content is, or what their motivation might be. Platforms like Facebook have unique visibility into such behavior when it takes place on our services, and are best suited to uncover and remove these surreptitious, often highly adversarial campaigns. This is why from 2017 through mid-2021, we have taken down and publicly reported on over 150 covert influence operations that violated our policy against Coordinated Inauthentic Behavior ("CIB"). They originated from over 50 countries worldwide and targeted both foreign and domestic public debate.

[1] In 2020, we began reporting on our broader enforcement against deceptive tactics that do not rise to the level of CIB, to keep adding to the public's understanding of these often financially-motivated behaviors. For more details on our efforts against inauthentic behavior, see our IB Report.

This threat report draws on our existing public disclosures and our internal threat analysis to do four things: *First*, it defines how CIB manifests on our platform and beyond; *Second*, it analyzes the latest adversarial trends; *Third*, it uses the US 2020 elections to examine how threat actors adapted in response to better detection and enforcement; and *Fourth*, it offers mitigation strategies that we've seen to be effective against IO.

While the defender community has made significant progress against IO, there's much more to do. Known threat actors will continue to adapt their techniques and develop new ones. To counter the evolving challenges to the integrity of public discourse — including domestic extremism and the increasingly blurry lines between speech and deceptive influence — we will need clear definitions and vigilance from across all of civil society.

Our hope is that this report will contribute to the ongoing work by the security community to protect public debate and deter covert IO.

Threat Trends

Since we published our first IO white paper in 2017, threat actors have continued to evolve their techniques. Here are some of the key trends and tactics we've observed:

1. **A shift from "wholesale" to "retail" IO**: Threat actors pivot from widespread, noisy deceptive campaigns to smaller, more targeted operations.
2. **Blurring of the lines between authentic public debate and manipulation**: Both foreign and domestic campaigns attempt to mimic authentic voices and co-opt real people into amplifying their operations.
3. **Perception Hacking**: Threat actors seek to capitalize on the public's fear of IO to create the false perception of widespread manipulation of electoral systems, even if there is no evidence.
4. **IO as a service:** Commercial actors offer their services to run influence operations both domestically and internationally, providing deniability to their customers and making IO available to a wider range of threat actors.
5. **Increased operational security:** Sophisticated IO actors have significantly improved their ability at hiding their identity, using technical obfuscation and witting and unwitting proxies.

6. **Platform diversification:** To evade detection and diversify risks, operations target multiple platforms (including smaller services) and the media, and rely on their own websites to carry on the campaign even when other parts of that campaign are shut down by any one company.

Mitigations

Influence operations target multiple platforms, and there are specific steps that the defender community, including platforms like ours, can take to make IO less effective, easier to detect, and more costly for adversaries.

1. **Combine automated detection and expert investigations:** Because expert investigations are hard to scale, it's important to combine them with automated detection systems that catch known inauthentic behaviors and threat actors. This in turn allows investigators to focus on the most sophisticated adversaries and emerging risks coming from yet unknown actors.
2. **Adversarial design:** In addition to stopping specific operations, platforms should keep improving their defenses to make the tactics that threat actors rely on less effective: for example, by improving automated detection of fake accounts. As part of this effort, we incorporate lessons from our CIB disruptions back into our products, and run red team exercises to better understand the evolution of the threat and prepare for highly-targeted civic events like elections.
3. **Whole-of-society response**: We know that influence operations are rarely confined to one medium. While each service only has visibility into activity on its own platform, all of us — including independent researchers, law enforcement and journalists — can connect the dots to better counter IO.
4. **Build deterrence.** One area where a whole-of-society approach is particularly impactful is in imposing costs on threat actors to deter adversarial behavior. For example, we aim to leverage public transparency and predictability in our enforcement to signal that we will expose the people behind IO when we find operations on our platform, and may ban them entirely. While platforms can take action within their boundaries, both societal norms and regulation against IO and deception, including when done by authentic voices, are critical to deterring abuse and protecting public debate.

Definitions

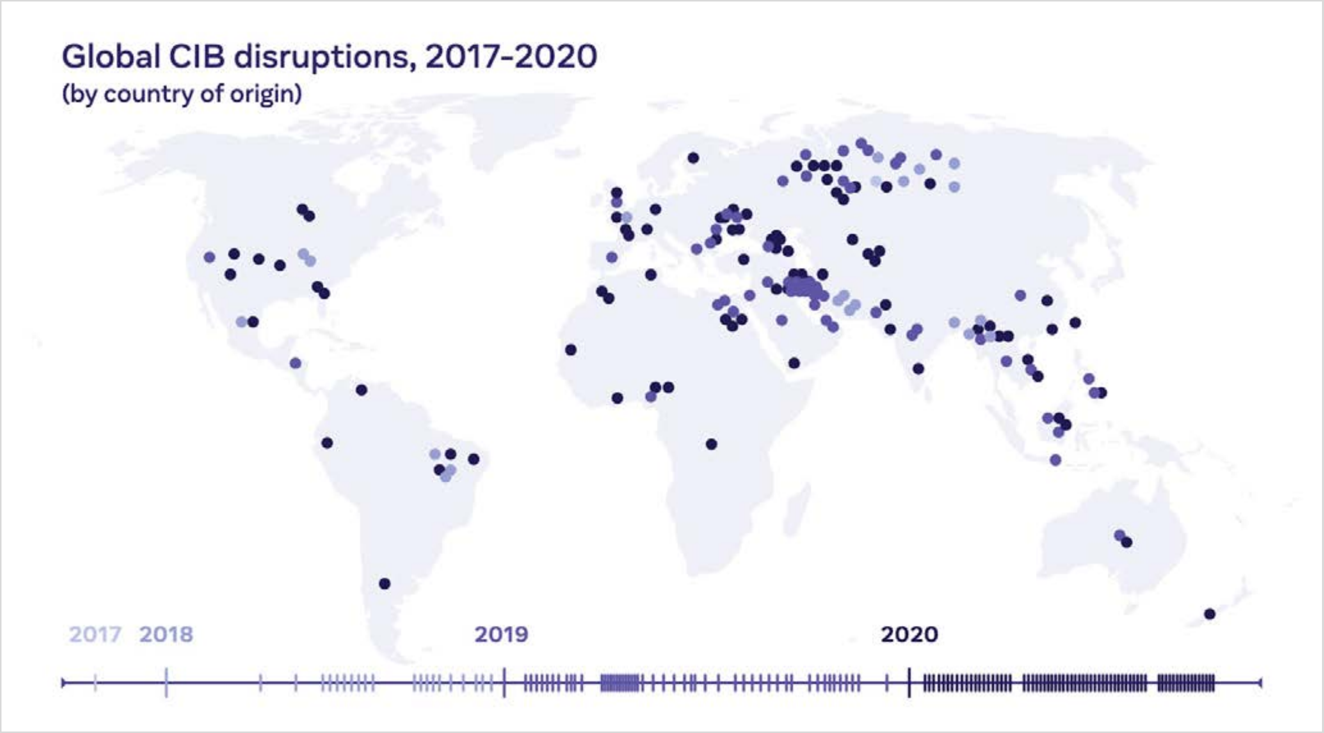
Influence Operations

Coordinated efforts to manipulate or corrupt public debate for a strategic goal.

Coordinated Inauthentic Behavior

A subset of influence operations, defined as "any coordinated network of accounts, Pages and Groups that centrally relies on fake accounts to mislead Facebook and people using our services about who is behind it and what they are doing."

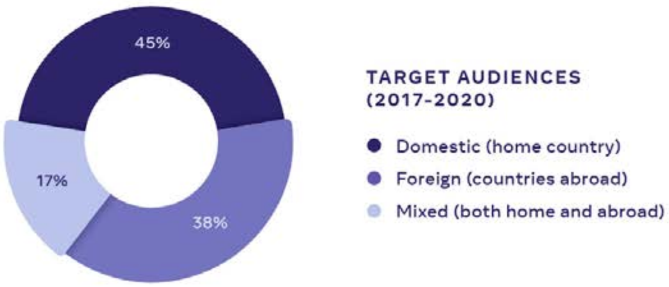
IO by the Numbers[2]



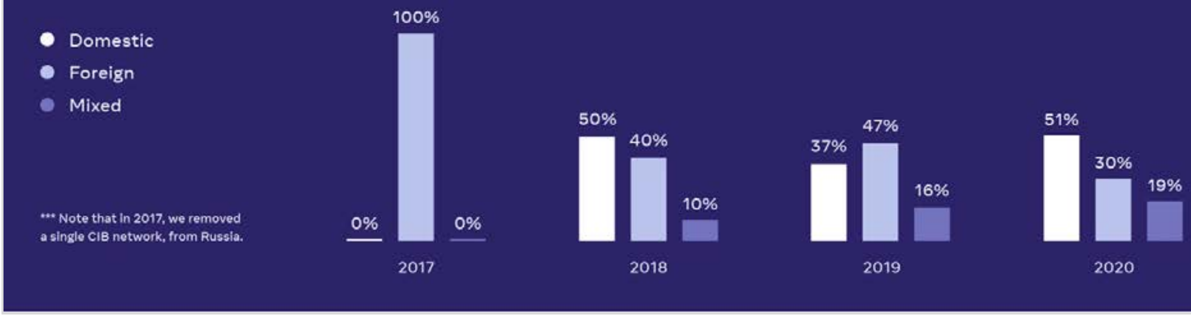
Global CIB disruptions, 2017-2020
(by country of origin)

[2] This report draws on over 150 CIB networks that we found and disrupted on our platform since 2017. Because our CIB reports share our findings with relative consistency, they provide a public record of threat evolution and response to known CIB networks. To the best of our knowledge, these reports constitute the most comprehensive record of both foreign and domestic IO operations, including state and non-state campaigns, and therefore provide a useful window into the global nature and trends of IO. These networks came from over 50 countries and operated in dozens of languages. We continue to grow our global capacity and will keep reporting our findings across various facets of influence operations.

Nature of Coordinated Inauthentic Behavior networks we disrupted



CHANGE IN NATURE AND TARGETING OVER TIME



*** Note that in 2017, we removed a single CIB network, from Russia.

Top 5 sources of CIB networks, 2017-2020 (countries of origin)

#1 Russia

27 CIB networks

Operators associated with (number of takedowns):

- IRA or Prigozhin's entities (15)
- Intelligence services (4)
- Media websites (2)

#2 Iran

23 CIB networks

Operators associated with (number of takedowns):

- Government including state broadcaster (9)

#3 Myanmar

9 CIB networks

Operators associated with (number of takedowns):

- Military or police (6)

#4 USA

9 CIB networks

Operators associated with (number of takedowns):

- Conspiratorial or fringe political actors (3)
- PR or consulting firms (2)
- Media websites (2)

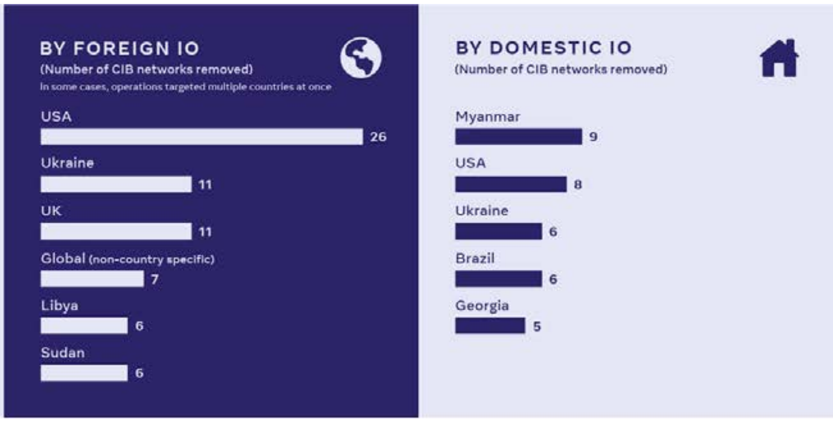
#5 Ukraine

8 CIB networks

Operators associated with (number of takedowns):

- PR & Ad agencies (3)
- Political parties (2)

Countries most frequently targeted by influence operations, 2017-2020



Introduction

Over the past four years, our security teams at Facebook have identified and removed over 150 networks for violating our policy against Coordinated Inauthentic Behavior ("CIB"). The CIB policy was a major piece of Facebook's broader security strategy against influence operations ("IO") developed in response to foreign interference by Russian actors in 2016. Since then, we've investigated and disrupted operations around the world, and these public enforcements offer a global picture of IO. These operations have targeted public debate across both established and emerging social media platforms, as well as everything from local blogs to major newspapers and magazines. They were foreign and domestic, run by governments, commercial entities, politicians, and conspiracy and fringe political groups.

Influence operations are not new, but over the past several years they have burst into global public consciousness. These campaigns attempt to undermine trust in civic institutions and corrupt public debate by exploiting the same digital tools that have diversified the online public square and empowered critical discussions from *Me Too* to the *Black Lives Matter* movements.

In response to this rising threat, a community of defenders that includes social media platforms, civil society advocates, open-source researchers, law enforcement, and the media have all fielded teams to expose IO and take it down. As part of this effort, our teams at Facebook built our own, blended enforcement strategy to not only detect and stop particular influence operations, but to expose the tactics behind them and make them less effective overall. With our partners, we have forced influence operators to work harder, only to get caught sooner. Still, there is more to be done.

The closing of the 2020 election season — where we saw a rapid evolution in adversarial behavior and the response to it — has created a valuable opportunity to look back at the lessons we've learned over the past four years, refine our collective understanding of the threats we face, and anchor the discussion going forward. This threat report is designed to help that effort, focusing on our enforcement against CIB.[3]

[3] In 2020, we also began publicly reporting on our broader enforcement against deceptive tactics that do not rise to the level of CIB to keep adding to the public's understanding of this adversarial space.

We will address four topics: *first*, we provide an update on our thinking about influence operations and how we define them today; s*econd*, we outline the adversarial threat trends that we've seen develop since our last report in 2017*; third*, through the prism of the US 2020 elections, we examine how threat actors adapted in response to better detection and enforcement; and *fourth*, we describe some of the more effective counter-IO techniques thus far and begin answering the question about what we can collectively do to further constrain future threats.

SECTION 1

Defining IO

Historically, influence operations have manifested in different forms: from covert campaigns that rely on fake identities to overt state media efforts that use authentic and influential voices to promote messages that may or may not be false on their face.

As we've studied emerging adversarial tactics and actors, our understanding of influence operations has evolved. Today, we define influence operations as:

> **Coordinated efforts to manipulate or corrupt public debate for a strategic goal.**

Tackling the many tactics that make up influence operations requires multiple approaches. Which is one reason why, when designing policies intended to counter IO, it is important to distinguish deceptive *content* from deceptive *behavior*.[4]

When someone posts deceptive *content* in their own name, platforms like Facebook can supply people with additional context about who they're hearing from, so they can validate the posts they're seeing (*e.g.* state media labels, voting information labels, fact-checking labels, *etc*.). If the content itself violates our Community Standards, we can remove it (*e.g. harmful health-related misinformation or voter suppression misinformation*) or reduce its distribution.

However, when an actor conceals their identity through deceptive *behavior*, the public can't judge who they are, how trustworthy their content is, or what their motivation might be. Platforms like Facebook have a unique visibility and are best suited to uncover and remove these surreptitious campaigns.

[4] Camille François, "Actors, Behaviors, Content: A Disinformation ABC," Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, September 20, 2019, https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf [Last accessed May 11, 2021]

The most egregious form of this type of deception is Coordinated Inauthentic Behavior (CIB):

> **CIB is any coordinated network of accounts, Pages and Groups on our platforms that centrally relies on fake accounts to mislead Facebook and people using our services about who is behind the operation and what they are doing.**

Since 2017, Facebook's Coordinated Inauthentic Behavior policy has been a primary vehicle for enforcing against these covert deception campaigns. For example, if someone or some entity creates a network of Facebook Pages designed to look like independent news organizations, but surreptitiously controls them using fake accounts on behalf of a political party, that's coordinated inauthentic behavior.

Our CIB definition has a distinct lower bound: it requires the central use of fake accounts to mislead people about who is behind the operation. This threshold ensures that we can be clear about what is and what isn't CIB, and that we can impose significant consequences against those who cross the line. Many deceptive efforts on our platforms don't cross the CIB threshold, and we address them through other policies.[5]

Because CIB is such a serious violation, when we find these operations, we remove all inauthentic and authentic accounts, Pages and Groups directly involved in this activity. When we enforce against CIB, we do so based on the deceptive *behavior* we see on our platform — not based on the content they share. We focus on behavior for two reasons.

First, content in and of itself isn't a reliable signal for determining whether a given account or a Page is part of an influence operation. We have seen deceptive campaigns reuse popular, authentic content to build an audience, as well as real people unwittingly post memes originally created by IO actors. Enforcing against influence operations based on content signals would have an overly broad impact on innocent people and innocuous posts.

Second, by enforcing based on behavior — regardless of who's behind the operation, their content or political bent — we maintain consistency worldwide and ensure the neutrality of our

[5] For inauthentic behavior that does not rise to the level of CIB, one tool we rely on is our Inauthentic Behavior policy. For more information, see our IB Report.

CIB policy application, even in the midst of critical civic moments.[6] This content-agnostic enforcement has been important because many influence operations focus on elections or major civic debates, leading to removals of operations, no matter which side they target or favor in public discussion.

This is also one of the reasons why we report our CIB takedowns publicly, describe the behavior we see, and share information with independent open-source researchers so they can form their own conclusions about the activity. Transparency helps create predictability and build trust in our enforcements by ensuring that others can review our actions. It also provides a consistent public record of our behavior-based enforcement.

Evolving threat landscape & response

Like any sufficiently complex societal phenomena, IO takes different shapes and forms and varies from platform to platform. As these campaigns evolve in response to better enforcement, they will continue to blur the boundary between legitimate advocacy and illegitimate manipulation.

Consider a few examples: Political campaigns have long paid canvassers to knock on doors, but when campaigns pay supporters or influencers to use fake or misleading online accounts to amplify their message on social media, does that cross a line into deception and manipulation?[7] Consider as well activists, governments or lobbyists who seek support for their causes by creating seemingly independent media entities to inject their message into the public discourse; or marketing firms amplifying particular narratives through Pages and Groups without disclosing who runs them. Such tactics exemplify the gray areas where the boundary between advocacy and deception can be hard to define, and pose important questions for how public debate should function online.

Over the past four years, threat actors have adapted their behavior and sought cover in the gray spaces between authentic and inauthentic engagement and political activity. We know they will continue to look for new ways to circumvent our defenses. This is why we are careful not to treat all IO as a singular problem. Instead, we rely on a broad definition of “influence

[6] To note, we ban particularly egregious violators, including entities that are primarily organized to conduct CIB, from our platforms. For example, the Russia-based Internet Research Agency and its affiliated entities, as well as a number of IO-for-hire services, have been banned for crossing this threshold.

[7] We will examine a particular network we disrupted that engaged in this behavior ahead of the US 2020 election in Section 3.4. “Operations originating from the US”

operations," and precisely define specific violations like "coordinated inauthentic behavior" under the IO umbrella.

We do so because it wouldn't be proportionate or effective to use the same policy to enforce against a foreign government creating fake accounts to influence an election in another country as we do against a political action committee that isn't fully transparent about the Pages it controls. Doing so would force us to either over-enforce against less serious violations, or under-enforce against the worst offenders.

We are continually refining the policy framework and enforcement tools we use to combat IO. But we also know that IO rarely operates on a single medium — threat actors target all of society including multiple platforms, traditional media and influential public figures. No one platform or institution can tackle this alone.

In this section, we shared the framework we use to understand and combat the IO activity we see on our platform. To counter IO effectively and holistically, we need to bring together perspectives from other platforms, civil society, government, and media. 2021 offers a golden opportunity to have this broader discussion, and we hope this paper informs that debate.

SECTION 2

The State of IO, 2017-2020

2.1 Threat actors

Over the past several years, much public attention has focused on “foreign interference” (*i.e.* covert foreign-origin influence operations) and the risk it poses to the integrity of elections and trust in democratic systems. While the most studied examples of contemporary IO were indeed run by foreign actors, influence operations are increasingly common tools for non-state and domestic actors.[8] Over the past several years, we have seen new actors emerge — including commercial entities and political interest groups — running both foreign and domestic IO campaigns.

It is also important to note that IO is not confined to efforts focused solely on elections. In fact, we have found long-running operations that focused on different topics at different times, ranging from elections to military conflicts and sporting events.[9]

Broadly speaking, we categorize influence operations based on both the type of actor behind them and the audience they target:

Actor

- **Government**: IO undertaken directly by state actors, including military, intelligence, and cabinet-level bodies.
- **Non-Government:** IO undertaken by groups unaffiliated with a government, including hacktivists, financially-motivated 'troll farms', commercial entities, political parties and campaigns, special interest or advocacy groups.

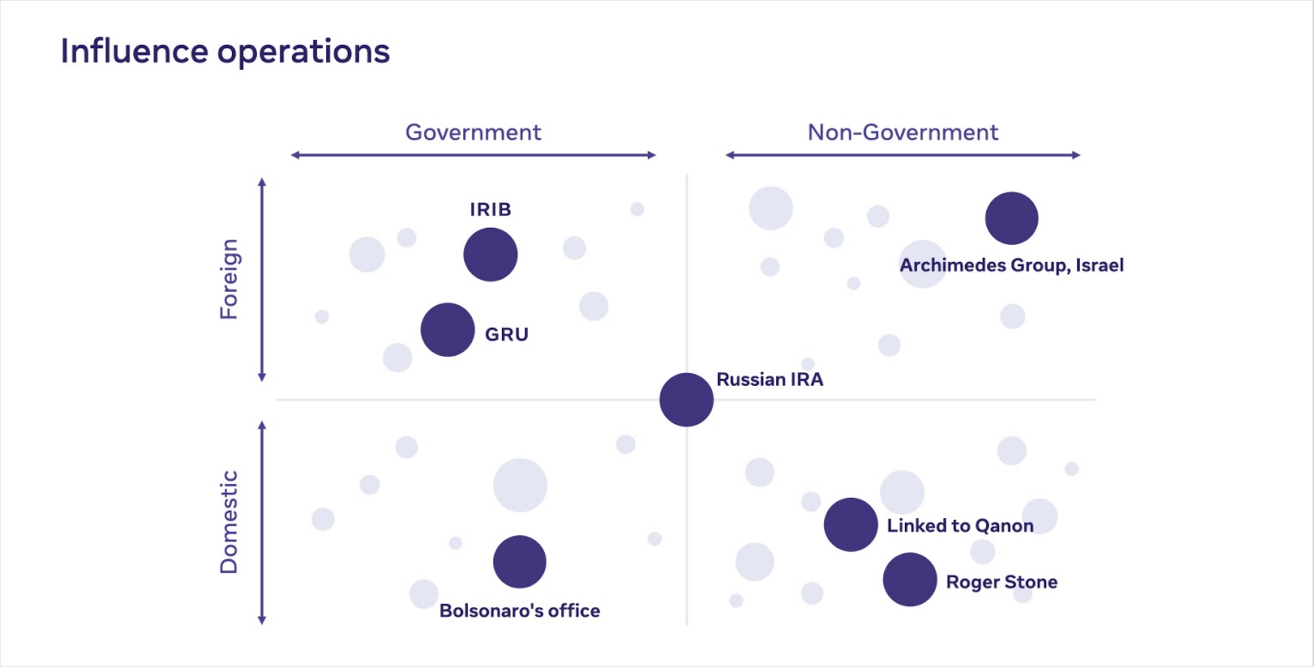
Target

- **Domestic:** IO that targets public debate in the same country from which it operates.
- **Foreign:** IO that targets the public debate in a different country from which it operates.
- **Mixed:** We also see IO campaigns and threat actors that run campaigns that target both domestic and foreign audiences.

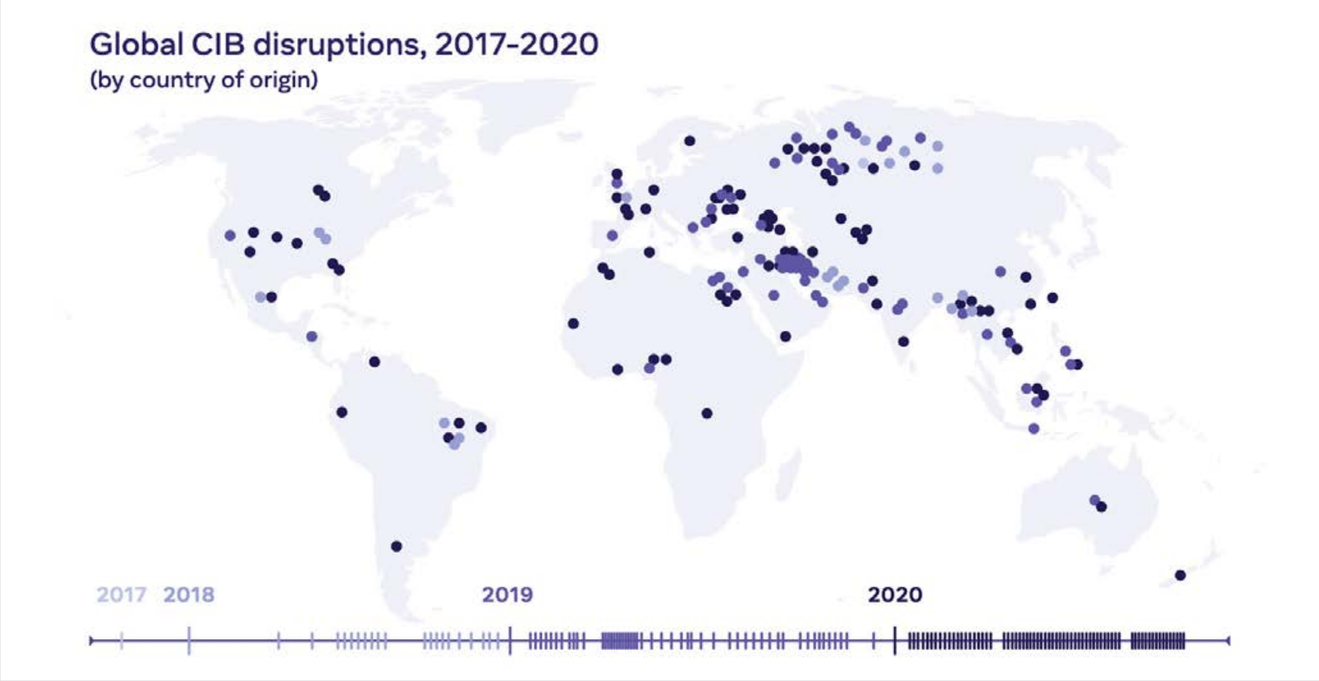
[8] About half of the CIB networks we took down over the years have been domestic in nature, *i.e.* they targeted audiences in the same countries they originated from. See Appendix 1.

[9] More details in Section 2.2. “Trends in IO”

The majority of influence operations that Facebook removed for CIB over the past four years tend to fall into multiple categories, along multiple axes, exemplifying the increasingly blurry lines and complex nature of this threat.

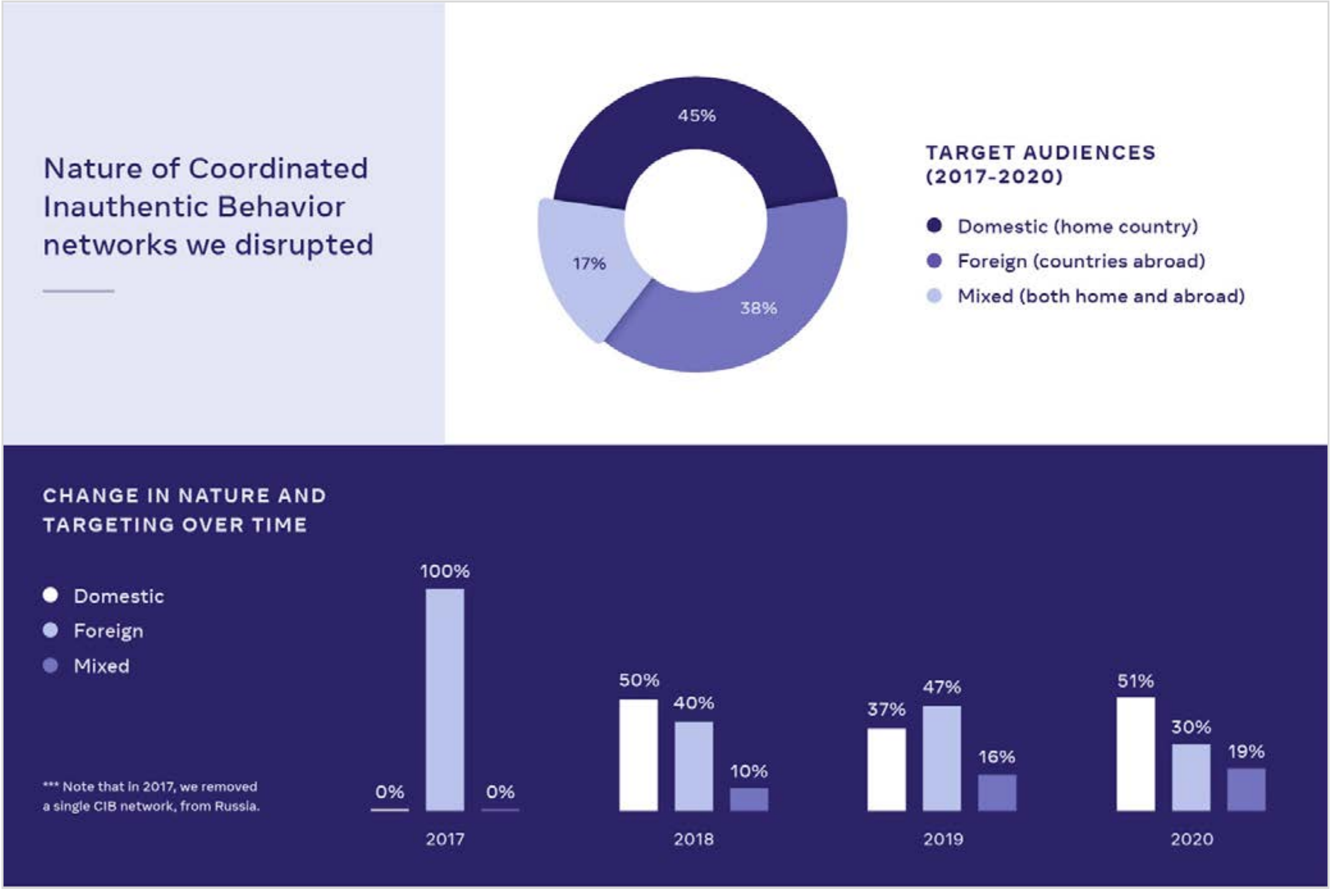


The following are a few high-level global snapshots of covert IO activity that we've found on our platform between 2017-2020[10]

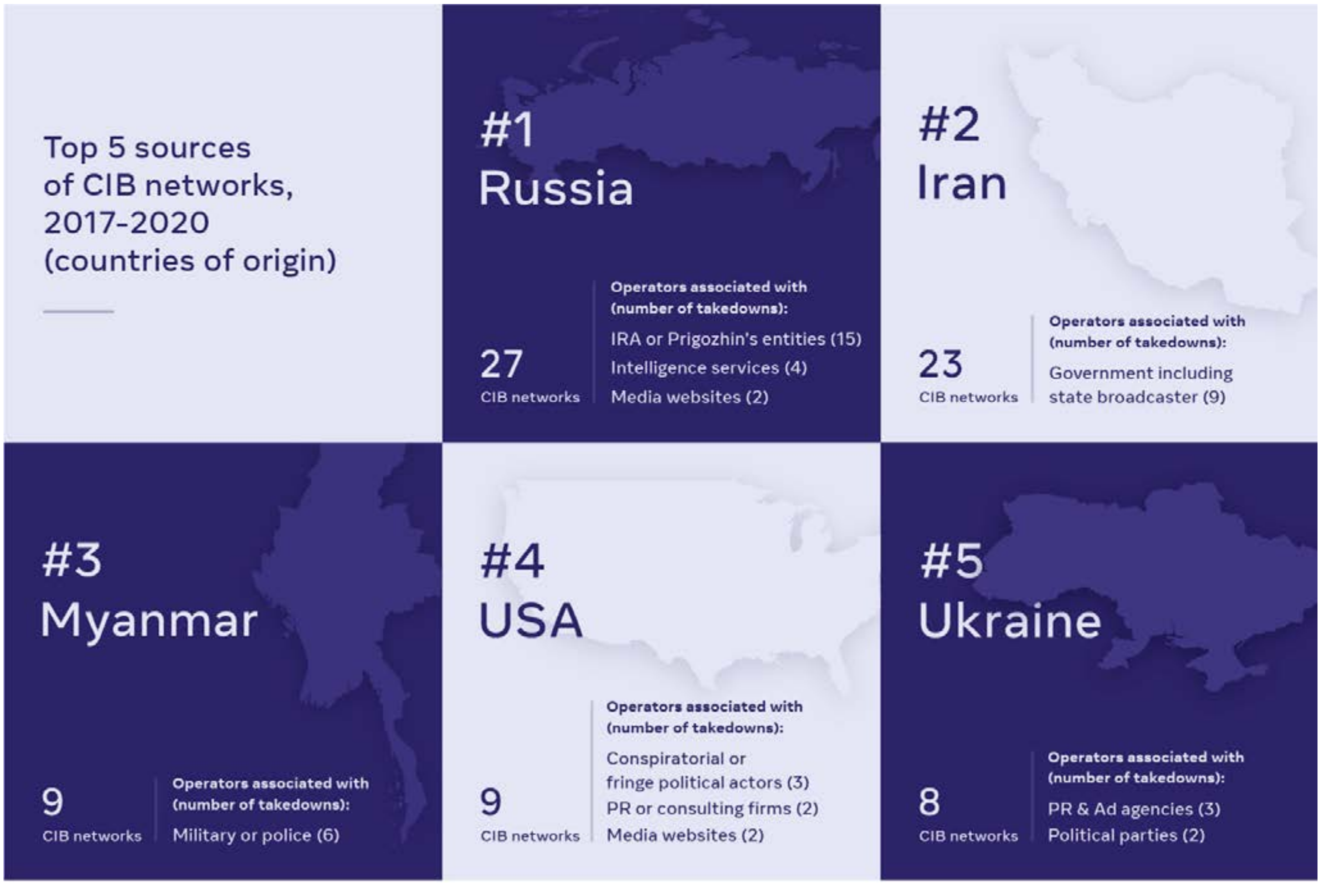


[10] To support further analysis of IO globally, in addition to our analysis, we are sharing a full list of Facebook's public disruptions that includes a set of summary statistics we have reported since September 2017. See Appendix 1

Of the more than 150 CIB operations we've taken down around the world to date, about half were domestic in nature, a slightly smaller portion focused solely on foreign countries, and the rest targeted audiences both at home and abroad. Despite the fact that public discourse in the US shifted from focusing on foreign operations in 2017-2019 to focusing on domestic operations in 2020, we continued to see significant portions of all three types, and a steady rise in mixed targeting from 2018 through 2020.



Here are the countries where most CIB networks we found on our platform came from:



Top 5 sources of CIB networks, 2017-2020 (countries of origin)

#1 Russia

27 CIB networks

Operators associated with (number of takedowns):
IRA or Prigozhin's entities (15)
Intelligence services (4)
Media websites (2)

#2 Iran

23 CIB networks

Operators associated with (number of takedowns):
Government including state broadcaster (9)

#3 Myanmar

9 CIB networks

Operators associated with (number of takedowns):
Military or police (6)

#4 USA

9 CIB networks

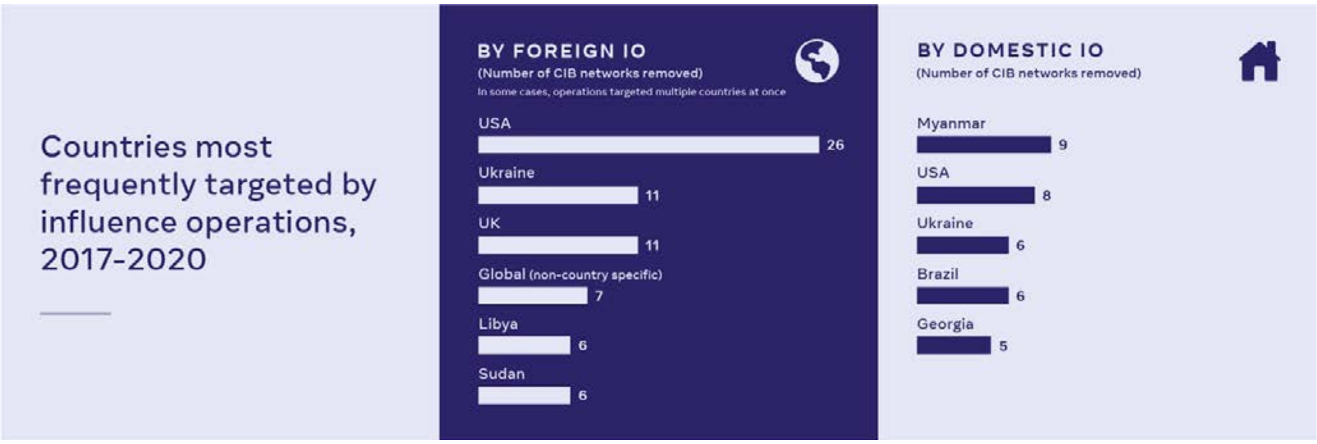
Operators associated with (number of takedowns):
Conspiratorial or fringe political actors (3)
PR or consulting firms (2)
Media websites (2)

#5 Ukraine

8 CIB networks

Operators associated with (number of takedowns):
PR & Ad agencies (3)
Political parties (2)

Finally, here is where domestic and foreign CIB networks that we found on our platform focused the most in the world:



Countries most frequently targeted by influence operations, 2017-2020

2.2. Trends in IO

In our 2017 report, we described the IO tactics and techniques we saw at the time, including threat actors' reliance on large numbers of fake accounts and the amplification of hacked and leaked information.

Since, we've seen change in threat actors' behavior in response to detection and enforcement across the internet. While the trends we highlight below are generally encouraging — the operations we've seen have moved to more targeted and often less effective techniques — we know that threat actors have not given up. As we collectively push IO actors away from easier venues of attack, we see them try harder to find other ways through. To counter these new attempts, we continue to stress-test and improve our defenses.

We've observed six key trends and tactics:

A shift from "wholesale" to "retail" IO

As we improved our automated blocking of fake accounts, we made it harder for threat actors to successfully operate high-volume, "wholesale" influence operations that broadcast messages to target audiences at a large scale. Some IO actors have since attempted to evade detection by shifting to narrower "retail" campaigns that use fewer assets and focus on narrowly targeted audiences.

For example, in May 2019 we removed an Iranian network that used a small number of fake accounts posing as journalists and other fictitious personas to seed and amplify their content. Rather than trying to broadcast it across social media, as earlier operations by Iranian actors had done, this campaign reached out directly to policymakers, reporters, academics, dissidents, and others. Their fictitious personas also submitted letters to the editor and wrote guest columns in U.S. newspapers, masqueraded as journalists soliciting interviews with politicians and pitched stories to reporters. Indeed, they succeeded in publishing their work in a number of legitimate publications, yet gained almost no following on Facebook.[11]

In another case from early 2020, we found and removed a network run by Russian military intelligence that focused on Ukraine and neighboring countries. They created fake personas

[11] Alice Revelli and Lee Foster, "Network of Social Media Accounts Impersonates U.S. Political Candidates, Leverages U.S. and Israeli Media in Support of Iranian Interests," FireEye, May 28, 2019, https://www.fireeye.com/blog/threat-research/2019/05/social-media-network-impersonates-us-political-candidates-supports-iranian-interests.html [Last accessed on May 11, 2021]

that operated across blogging forums and multiple social media platforms. Some of them posed as citizen journalists and tried to contact policymakers, journalists and other public figures in the region. As with the Iranian network above, this operation had not built a large following on Facebook when we removed it, but a few of its blogs were picked up by outlets not run by the operation.[12]

Each fake account in these "retail" campaigns takes more time and effort to create because the actors invest heavily in developing more credible online personas so they can't be as easily spotted. This includes creating fake personas that span multiple platforms as evidentiary "backstops" for when researchers or journalists or the public try to verify their identity. By building out these fictitious personas to appear more legitimate across multiple services, these operations attempt to mislead the public and evade detection and removal by Facebook and other platforms.

Still, despite their relatively sophisticated nature, both of these operations reveal one of the fundamental challenges of "retail" IO — without a lucky break, they go nowhere. Neither the Iranian nor the Russian operation gained significant traction or attention.

Blurring the lines between authentic public debate and deception

Foreign operations

When "retail" operations are run by foreign actors, they typically attempt to mimic the domestic, authentic audiences they target so they can more credibly exploit contentious political and societal issues in a given country.

Particularly sophisticated foreign actors are getting better at blurring the lines between foreign and domestic activity by co-opting unwitting (but sympathetic) domestic groups to amplify their narratives. They may also attempt to purchase compromised assets or directly compromise domestic actors through social engineering or hacking to gain access to their built-in audiences. These already-established communities then become unsuspecting but active amplifiers of IO campaigns.

Such convergence makes it more difficult to distinguish illegitimate manipulation efforts from legitimate civic discourse, and poses a challenge in delineating the appropriate scope of enforcement.

[12] Ben Nimmo, Camille François, C. Shawn Eib and L. Tamora, "From Russia With Blogs," Graphika, February 12, 2020, https://graphika.com/reports/from-russia-with-blogs/ [Last accessed on May 11, 2021]

For example, in July 2018, we removed a network linked to the Russian Internet Research Agency ("IRA") that was engaging with pre-planned, authentic events. They would target events focused on particularly hot-button issues and volunteer to amplify them on behalf of the local organizers. Interestingly, when the IRA attempted to create its own events on Facebook in early 2016, they often failed to gain traction as they were unable to build an audience without the reach of authentic local groups.

Of course, this approach is not without its own risks, particularly in countries with a robust civil society and democratic governments determined to root out IO. When platforms and law enforcement are on the lookout, threat actors who attempt to outsource manipulation to locals increase the ways in which they can be detected.[13] We saw this in particular ahead of the US 2020 election, which we'll cover in more detail in Section 3: Targeting the US ahead of the 2020 election.

Domestic operations

Domestic IO also continues to push the boundaries of acceptable online behavior worldwide. About half of the influence operations we've removed since 2017 – including in Moldova, Honduras, Romania, UK, US, Brazil and India – were conducted by locals that were familiar with domestic issues and audiences. These were political campaigns, parties, and private firms who leveraged deceptive tactics in the pursuit of their goals.

For example, in 2018, we removed accounts operated by New Knowledge, a US firm that engaged in misleading tactics during the 2017 Alabama special election. In particular, they created a Facebook Page posing as conservative Alabamians that, among other things, attempted to steer conservatives towards a write-in candidate. We will share more examples in Section 3.

We anticipate seeing more local actors worldwide attempt to use IO tactics to influence public debate in their own countries, further blurring the lines between authentic public debate and deception. In turn, technology platforms, traditional media and civil society will be faced with more challenging policy and enforcement choices.

[13] See for example: Michael Schwirtz and Sheera Frenkel, "In Ukraine, Russia Tests a New Facebook Tactic in Election Tampering," New York Times, March 29, 2019,
https://www.nytimes.com/2019/03/29/world/europe/ukraine-russia-election-tampering-propaganda.html
[Last accessed on May 11, 2021]

Going forward, it will be increasingly critical that we as a society proactively engage in a broader discussion on what constitutes acceptable online political behavior and what doesn't, and how we can distinguish between public diplomacy and influence operations, political campaigning and election manipulation, political activism and engagement hacking.

There isn't always a clear line between authentic and deceptive tactics in this space, and we should collectively determine how to tackle this challenge without encroaching on free speech and other democratic values.

The emergence of perception hacking

As it has become more difficult to run large covert influence operations on social media, some IO actors have engaged in what we call "perception hacking." That is, rather than running actual on-platform campaigns or compromising election systems, they are attempting to garner influence by fostering the *perception* that they are everywhere, playing on people's fear of widespread deception itself.

For example, in the waning hours of the 2018 US Midterm elections, we investigated an operation by the Russian Internet Research Agency that claimed they were running thousands of fake accounts with the capacity to sway the election results across the United States. They even created a website — usaira[.]ru — complete with an "election countdown" timer where they offered up as evidence of their claim nearly a hundred recently-created Instagram accounts. These fake accounts were hardly the hallmark of a sophisticated operation, rather they were an attempt to create the *perception of* influence.

By the time this list appeared online however, in collaboration with law enforcement and industry peers, we had already investigated and removed a small network of accounts on Facebook and Instagram originating in Russia, including those listed on this website. We also determined that there was no evidence of a larger operation, and the IRA's broader claims were false.

By sharing this context with press and civil society experts, we were able to stymie this avenue of IO without amplifying the underlying manipulation efforts or giving the threat actors the very attention they were seeking in the first place. While disclosing IO in the midst of the election campaign can be challenging, this case cemented the importance of being public about our findings in tackling perception hacking.

We saw similar tactics used by Iranian and Russian actors ahead of the US 2020 election, which we'll cover in more detail in Section 3.

By intermingling minimally authentic information with falsified data or claims, perception hackers try to take advantage of the highly competitive and time-pressured media and political environments to pollute public discourse and exploit existing societal divisions and uncertainty. Ultimately, they try to force the defender community to prove a negative (*i.e.* that IO does not exist), which, even if successful, could create more "noise" and uncertainty, further seeding distrust in public institutions like elections.

Enforcing amidst this "noise" can be challenging, particularly under time constraints ahead of major elections. Some IO actors may also seek to benefit from their unsuccessful operations getting caught and publicly exposed because it still allows them to create some uncertainty about the information environment and tout their alleged impact.

While it's a challenging balance to strike, we work to mitigate risk through consistent public reporting, no matter where the activity is coming from or who's behind it. This creates a default expectation of exposure, and hopefully helps to reassure the public over time that although these operations occur, the fact that they were removed doesn't mean that they were necessarily successful.

As we saw ahead of the US 2020 Presidential election, collaboration among industry, government and civil society can be effective in countering this tactic and inoculating public debate against IO. We will discuss some of the proven measures against IO efforts like these in Section 4: "Countering IO."

The rise of IO-for-hire

Over the past four years, we have investigated and removed influence operations conducted by commercial actors — media, marketing and public relations companies, including in Myanmar, the US, the Philippines, Ukraine, the UAE and Egypt. Some of these operations were domestic, promoting interests aligned with political entities from within their country of origin. Some offered IO services to paying clients both at home and abroad, making these techniques accessible to those with less resources or infrastructure to run their own IO campaigns.

These commercial, “IO-for-hire” entities could also be used by sophisticated actors to hide their involvement behind private firms, making attribution more challenging. Of course, from a strategic perspective, IO-for-hire is not without its risks for influence operators.

Notably, because these commercial firms operate across multiple regions, their content may lack the necessary domestic context to be convincing, which makes it more difficult to gain a following or go unnoticed by platforms and local civil society.

In May 2019, for example, we identified and removed an Israeli firm — Archimedes Group — that was running campaigns on behalf of its clients in Nigeria, Senegal, Togo, Angola, Niger and Tunisia, along with some activity in Latin America and Southeast Asia. This network repeatedly made blatant mistakes in their posts regarding the on-the-ground reality in the countries they targeted.[14]

It takes significant resources and time to build the kind of business, brand and infrastructure necessary to run effective influence operations across multiple platforms. Moreover, these entities are at constant risk of being detected by platforms and domestic law enforcement. When their operations are discovered, these companies lose their on-platform assets and in the most severe cases they are banned from ever coming back to our platform. Even if they start over and try harder to hide, this does not make for a sustainable business model in the long run.

We’ve seen commercial IO vary in sophistication, with many still relying on outdated tactics that Facebook and other defenders have gotten better at detecting. Some of these measures, including our Page transparency tools, have enabled researchers, investigative journalists and the public to see who’s behind the Pages and ads they interact with on Facebook, and in some cases find and flag suspicious activity to us so we can investigate and take action.

In addressing this particular trend, we continue to disrupt this emerging business model, including through building deterrence so that IO actors will incur reputational, legal and financial costs when their activity is found.[15]

[14] Luiza Bandeira, Andy Carvin, Kanishk Karan, Mohamed Kassab, Ayushman Kaul, Ben Nimmo and Michael Sheldon, “Inauthentic Israeli Facebook Assets Target the World,” DFRLab, May 17, 2019, https://medium.com/dfrlab/inauthentic-israeli-facebook-assets-target-the-world-281ad7254264

[15] More detail on what defense measure worked and what didn’t can be found in Section 4 “Countering IO”

Increased operational security

In response to increased efforts at stopping them, the more sophisticated threat actors — including from Russia and China — have improved their operational security ("OpSec"). They are showing more discipline to avoid careless mistakes like logging into purportedly American accounts from St. Petersburg in Russia. Some are also getting better at avoiding language discrepancies and similar inauthenticity clues by re-appropriating authentic content from domestic communities, rather than creating their own.

For example, in October 2019, we removed a Russian IRA-linked network that was among the first to target the US 2020 election. It primarily posted other people's content, including memes with minimal or no text in English, and screenshots of social media posts by news organizations and public figures. In addition to this front-end obfuscation, the campaign had the hallmarks of a well-resourced operation that took consistent OpSec steps to conceal their identity and location on the back-end.

"Secondary Infektion," another Russia-linked network that spanned over 300 platforms and services, currently remains unattributed, beyond its geographic origin.[16] Our team was the first to expose this activity in May 2019, starting off a series of disruptions across many platforms and independent investigative research by journalists and IO experts. Of hundreds of this network's separate attempts to inject its narratives into mainstream conversations, only one story managed to break through, and only after it was amplified by one of the top political figures in the UK, ahead of the 2019 election.[17]

Research into this operation and others like it highlights an interesting feature: better OpSec in influence operations comes with significant trade-offs around engagement. As we've seen repeatedly, using one-time-use or 'burner' accounts makes it difficult to gain followers or have people see your posts at all. Consistently hiding who you are and only resorting to copying other people's existing content fails to build a distinct voice among authentic communities.

Going forward, we expect these tactics will continue to evolve in response to enforcement measures across technology platforms. It will continue to be critical that political and public

[16] Ben Nimmo, Camille François, C. Shawn Eib, Lea Ronzaud, Rodrigo Ferreira, Chris Hernon, and Tim Kostelancik, "Exposing Secondary Infektion," Graphika, June 16, 2020, https://www.graphika.com/reports/exposing-secondary-infektion/ [Last accessed on May 11, 2021]

[17] Jack Stubbs, "Leak of papers before UK election raises 'spectre of foreign influence' - experts," Reuters, December 2, 2019, https://www.reuters.com/article/uk-britain-election-foreign-idUKKBN1Y6206 [Last accessed on May 25, 2021]

figures, our colleagues in the IO research community and journalists remain vigilant against the attempts to amplify malicious manipulation campaigns.

Platform diversification

Likely in response to intensified detection and enforcement against IO, we saw a shift to operations that target multiple platforms — both online and off. We've seen this from both experienced threat actors and newcomers in the IO space. By running operations on multiple platforms, threat actors are likely trying to ensure that their efforts survive enforcement by any given platform. They've also targeted hyper-local platforms (*e.g.* local blogs and newspapers), to reach specific audiences and to target public-facing spaces with less resourced security systems.

For example, in February 2020, we removed a network operated by an Indian digital marketing firm, aRep Global. It focused on a wide range of topics: from politics in the Gulf region to the 2022 FIFA World Cup in Qatar. This operation attempted to drive people to their websites posing as news outlets and relied on nearly a dozen platforms including Facebook, Instagram, Twitter, YouTube, Reddit, and Medium.[18]

However, as a number of recent cross-industry investigations have demonstrated, this approach has its drawbacks. Seeding stories on hyper-local blogs by fictitious characters in an attempt to attract authentic engagement further down the road may help create additional layers of obfuscation. But it also makes running IO more resource-intensive and leaves footprints across many more surfaces.

Managing these assets across multiple services means developing numerous backstops so that these fake entities can withstand scrutiny by platforms, law enforcement, researchers and journalists. To date, we have not seen successful examples of this approach run by automation. This strategy also increasingly resembles what intelligence-led influence operations looked like in the past, before the internet: highly targeted, expensive, and often of limited scope and impact.

To counter these campaigns, we have worked closely with our counterparts at tech companies and across civil society, and this effort should continue to expand going forward to study these

[18] Ben Nimmo, Camille Francois, C. Shawn Eib and L. Tamora, "Operation Red Card," Graphika, March 2, 2020, https://graphika.com/reports/operation-red-card/ [Last accessed on May 11, 2021]

networks' cross-platform behaviors. We've seen a number of operations disrupted often and early due to collaboration among our industry peers and researchers.

One example involved an Iranian network linked to the Islamic Republic of Iran Broadcasting Corporation. Starting with a 2018 Facebook investigation and takedown in collaboration with the cybersecurity company FireEye, this operation saw multiple waves of enforcement across platforms. Each subsequent takedown further shrank this network's ability to reconstitute itself and gain traction. Part of this success is due to a growing knowledge base among the defender community about these repeat offender networks. The more we observe and share about their behaviors and the technical signals associated with their activities, the more successful we all become in detecting them earlier in their life cycle.

SECTION 3

Case Study: Targeting the US Ahead of the 2020 Election

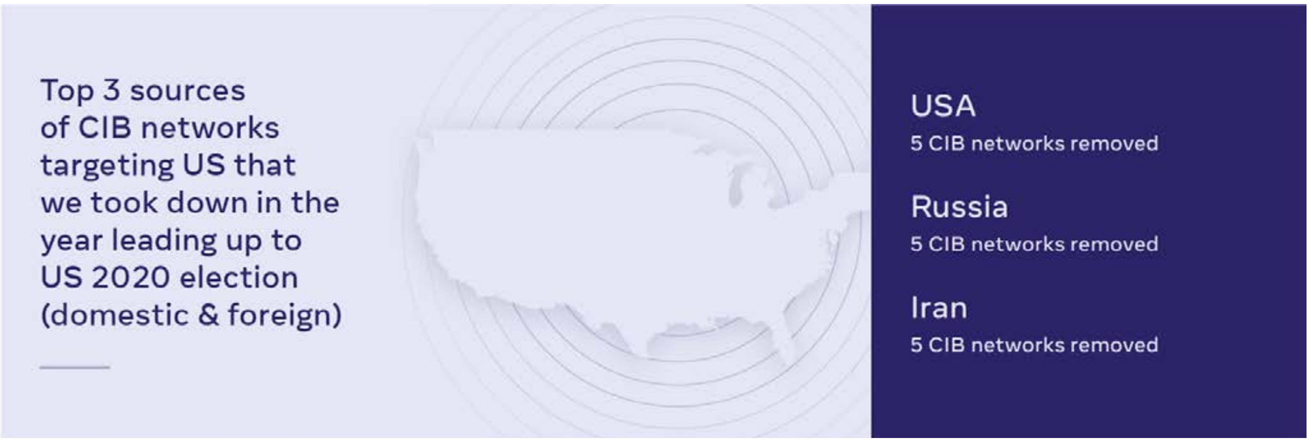
The 2016 US presidential election was a watershed moment in the recent history of influence operations. It triggered a global policy debate and similarly global response among the technology industry, governments and civil society. Throughout the 2016, 2018, and 2020 US elections, we've seen both some of the more active adversarial adaptation and also the response from the defender community, which made it a good candidate for a brief case study to examine what worked and what didn't.[19]

In this section, as a case study, we will detail our findings and actions taken against networks targeting the US in the year leading up to the 2020 US presidential election. We will highlight several notable changes we've seen since the 2016 election and raise questions about emerging risks to the information environment moving forward.

In the year leading up to the US 2020 election, we exposed over a dozen CIB operations targeting US audiences, including an equal number of networks originating from Russia, Iran, and the United States itself (we also share how the activity from China appeared very differently on our platform).

Some of these operations referred to the election explicitly, while others focused on general political and civic commentary and appeared to be in audience-building mode at the time we disrupted them. Some were run by repeat offenders who targeted earlier US election cycles, while others came from new or unknown actors. All of them attempted to interfere with public discourse by targeting American audiences in an election year using networks of inauthentic assets.[20]

[19] It's important to note that although this section of the report is focused on IO that targeted the United States in the year leading up to the 2020 election, the report more broadly confirms that influence operations are truly a global phenomenon. The campaigns that we have taken down since 2017 originated in over 50 countries, with the majority coming from or focused outside the US.
[20] More information on these operations can be found in our full list of disruptions in Appendix 1.



Top 3 sources of CIB networks targeting US that we took down in the year leading up to US 2020 election (domestic & foreign)

USA
5 CIB networks removed

Russia
5 CIB networks removed

Iran
5 CIB networks removed

Operations originating from Russia

We identified one CIB network linked to Russian military intelligence, which focused primarily on countries in Russia's immediate neighborhood, with the US as only a minor focus. However, the majority of the CIB takedowns from Russia came from IRA-linked actors. Even though these operations targeted a range of countries, many focused on the United States as the core of their activity. In addition to these US-focused campaigns, we also found and removed several operations that didn't target the US at the time of disruption, but were linked to actors associated with the 2016 election interference in the US.

Many of the Russia-origin operations exhibited the broader trends discussed in Section 2, including co-opting of authentic voices, a pivot to "retail" influence, and an increase in operational security to evade detection. Most interestingly, a number of IRA-associated campaigns actively recruited unwitting people, including activists and journalists, to write their content, manage their accounts, and try to circumvent our restrictions on posting political ads. This is a good example of the ongoing shift of IRA-linked operations from running their own large campaigns to using cutouts, smaller networks and their own websites, likely in response to detection and repeat removals.

One of these operations employed people in Ghana to focus on racial equality in the US. The campaign relied heavily on authentic accounts and off-platform coordination, including setting up an office in Accra for a fictitious NGO. We assessed that the operation outsourced this activity in an effort to appear more credible and authentic, minimize language discrepancies, and frustrate our ability to attribute. While we began this investigation internally, our collaboration with investigative journalists at CNN and Twitter was critical in understanding the on-the-ground operations behind this network.[21] We disrupted it before it could gain a meaningful following in the US.

Another IRA-linked effort set up two websites posing as news outlets at opposite ends of the political spectrum (Peacedata[.]net and NAEBC[.]com). To appear more legitimate, IRA operators created sophisticated fake personas with profiles on multiple platforms claiming to be the editors of these sites. They recruited freelance journalists, including people in Europe and America, to write on social and political issues targeting both the right and the left. While the left-leaning Peacedata network had a short and small presence on Facebook, NAEBC's attempt to create a fake account was detected and blocked by our automated systems before we'd even begun our investigation. Just as in the Ghana case, this operation attempted to run political ads, including by co-opting people in the US to do so.[22]

Finally, we identified a network targeting the US and operated by individuals in Mexico. They posted in Spanish and English about topics like feminism, Hispanic identity and pride, and the Black Lives Matter movement. Some of these fake accounts claimed to be associated with a nonexistent marketing firm in Poland. Others posed as Americans supporting various social and political causes and tried to contact real people to amplify their content. As expected, improved operational security made it challenging to determine who was behind this operation using on-platform evidence. In fact, we did not see sufficient evidence to conclusively attribute this operation beyond the individuals in Mexico who were directly involved. However, following our public disclosure of the operation, the FBI further attributed the activity to Russia's Internet Research Agency.[23]

Operations originating from Iran

Among the five Iranian IO operations that were identified to be focused on the US, two were linked to individuals associated with the Iranian government and its state broadcaster, IRIB. Notably, for the first time, we saw Iran-based actors engage in 'perception hacking,' which was one of the threats we were particularly concerned about coming from Russia in the lead-up to the US 2020 election.

[21] Clarissa Ward, Katie Polglase, Sebastian Shukla, Gianluca Mezzofiore and Tim Lister, "Russian election meddling is back - via Ghana and Nigeria - and in your feeds," CNN, April 11, 2020, https://edition.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html [Last accessed on May 11, 2021]

[22] Adam Rawnsley, "She Was Tricked by Russian Trolls—and It Derailed Her Life," The Daily Beast, September 6, 2020, https://www.thedailybeast.com/she-was-tricked-by-russian-trollsand-it-derailed-her-life. [Last accessed on May 21, 2021]

[23] "Foreign Threats to the 2020 US Federal Elections," National Intelligence Council, March 16, 2021, https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf.

In a shift from social media-based campaigns, about a week before the US vote, Iranian actors attempted a primarily email-based campaign posing as the Proud Boys, a US-based hate group.[24] They claimed to have compromised the US voting systems and threatened people to vote a certain way. Based on a tip from the FBI, we investigated and removed a single fake account created just days earlier in October 2020 by Iranian actors linked to the government in an attempt to seed these false claims. Their attempt failed to gain traction, and it was quickly exposed by law enforcement and tech platforms, as part of the ongoing pre-election collaboration to stop influence operations targeting the vote.

Operations originating from China

The China-origin activity on our platform manifested very differently than IO from other foreign actors, and the vast majority of it did not constitute CIB. Much of it was strategic communication using *overt* state-affiliated channels (*e.g.* state-controlled media, official diplomatic accounts) or large-scale spam activity that included primarily lifestyle or celebrity clickbait and also some news and political content.[25] These spam clusters operated across multiple platforms, gained nearly no authentic traction on Facebook, and were consistently taken down by automation.

We identified one China-originating covert network in September 2020 which took consistent operational security steps to conceal their location. It was operated from the Fujian province of China and focused primarily on Southeast Asia and on maritime security in the Asia-Pacific region. The United States was only a minor target of the operation. The few assets that focused on US politics claimed to support politicians from both major parties, but struggled to engage authentic users and build a significant audience.

Operations originating from the US

Ahead of the November election, we took action against a number of domestic CIB networks in the United States. Only some of them specifically targeted political conversation around the election. More than half were campaigns operated by conspiratorial and fringe political actors that used fake accounts to amplify their views and to make them appear more popular than they were. In addition to removing these deceptive networks, other teams across Facebook

[24] It is of note that this operation used email as its primary delivery channel, with social media playing a secondary role. This illustrates both the cross-platform nature of recent IO, and the shift from "wholesale" to "retail" approach, described in section 2.2.1.

[25] Ben Nimmo, Camille Francois, C. Shawn Eib and Lea Ronzaud, "Spamouflage Goes To America," Graphika, August 12, 2020, https://graphika.com/reports/spamouflage-dragon-goes-to-america/ [Last accessed May 11, 2021]

also worked to disrupt white supremacy, militia and conspiracy groups who spoke with their own voice yet engaged in aggressive, adversarial adaptation against our enforcement.

Most notably, one of the CIB networks we found was operated by Rally Forge, a US-based marketing firm, working on behalf of its clients including the Political Action Committee Turning Point USA. This campaign leveraged authentic communities and recruited a staff of teenagers to run fake and duplicate accounts posing as unaffiliated voters to comment on news Pages and Pages of political actors.[26] Its election-focused behavior began in the run-up to the 2018 midterms; it then went largely dormant until June 2020. In 2018, some of the accounts posed as left-leaning individuals to comment on content. Their later activity included creating what we call "thinly veiled personas" whose names were slight variations on the names of the people behind them, and which were solely dedicated to this deceptive campaign. We believe this shift in tactics was likely due to the majority of this network's fake accounts getting caught by our automated detection systems.

This particular case raises questions about the boundaries between acceptable political discourse and abusive deception. While in the physical world it is not uncommon to pay people to knock on doors and advocate for a particular position, the implications and potential harm are very different when people are hired to do the same using fake accounts online.

The US 2020 election campaign brought to the forefront the complexity of separating bad actors behind covert influence operations from unwitting people they co-opt or domestic influencers whose interests may align with threat actors. We found and removed IO attempting to get authentic voices to post on their behalf. We also saw authentic voices, including the then-US President, promoting false information amplified by IO from various countries including Russia and Iran.

This convergence makes it extremely complex for any one platform, government agency or media entity to counter IO. While we've seen successful attempts at debunking misleading claims by platforms, government agencies and newsrooms, these measures can't entirely short-circuit the spread of misleading information throughout the information environment when shared by influential authentic voices and reported on by the media.

[26] Isaac Stanley-Becker, "Pro-Trump youth group enlists teens in secretive campaign likened to a 'troll farm,' prompting rebuke by Facebook and Twitter," Washington Post, September 15, 2020, https://www.washingtonpost.com/politics/turning-point-teens-disinformation-trump/2020/09/15/c84091ae-f20a-11ea-b796-2dd09962649c_story.html. [Last accessed on May 21, 2021]

There is more to do and these learnings should be taken as an opportunity to evolve our society-wide defenses against IO. We share some of our own recommendations in Section 5 “Conclusion.”

SECTION 4

Countering IO

As we discussed in earlier sections, influence operations target multiple platforms and segments of public debate. There are specific steps we all can take to make IO less effective and easier to catch, and to help prepare the public to resist the operations that do make it through. This section will provide examples of how we think about these parallel efforts.

4.1. Combination of automation and expert investigations to remove IO

Over the past four years, we have taken down IO networks from over 50 countries. We've built a cross-disciplinary team focused on finding and disrupting sophisticated influence operations and then using insights from these investigations to improve our automated detection and enforcement at scale.

The team leading this effort has grown to over 200 people, whose expertise ranges from open-source research, threat investigations, cyber security, law enforcement and national security, to investigative journalism, engineering, data science and academic studies in disinformation.

The reason why we rely on such a diverse set of skills is to create scalable responses to IO, rather than focus on one-off CIB takedowns. Our goal is to build a fast and responsive product and security innovation cycle where we translate what we learn from investigations into product design, automated detection, policy development, and threat modeling. It enables us to continue improving our defenses in response to adversarial adaptation.

Here is how it works in practice:

Facebook's IO Threat Intelligence Team focuses on uncovering and understanding high-fidelity signals from the most sophisticated networks we disrupt. We know that this expert work is hard to scale. That's why, through our investigations, we identify behaviors and technical "signatures" that are common for a particular threat actor, as well as some tactics that are common across multiple influence operations. We then work to automate detection of these techniques at scale, in addition to modifying our products to make those behaviors more difficult and costly to bad actors and more transparent to users. This in turn compels threat actors to constantly adapt while our investigators can focus on finding new threats.

This approach also helps us build on our prior investigations because we know that threat actors will try to come back. After removing each CIB network, we keep monitoring for attempts by these actors to re-establish presence on our platforms. When they do, we take them down using both automation and manual detection, which allows us to scale our defense.

As we learn more about a particular threat actor, we develop a deeper understanding of their tactics and how they operate. This lets us not only find operations earlier in their life cycle, but also helps us to go back and apply these learnings to find older inactive assets-linked to them that we might have missed. For example, in April 2020, we took down a network linked to the Islamic Republic of Iran Broadcasting Corporation. Based on the knowledge we gained from several earlier investigations beginning in 2018, we were able to attribute and remove clusters of activity in multiple languages, some of which had been inactive since as far back as 2012.

In addition to proactively detecting malicious activity by known threat groups, we take steps to ensure that we aren't missing new techniques from yet unknown emerging actors. As part of that effort, we run threat ideation exercises to identify new risks and also actively collaborate with security teams at other tech platforms, independent researchers and government partners so we continue to improve our understanding of the threat environment.

All in all, this combination of automated and expert investigative detection makes it difficult for these campaigns to remain active and undetected for long periods of time.

Fake account enforcement is a good example of this work. Because we know that fake accounts are often one of the core elements of IO, we are continually improving our machine learning systems that help us automate detection. We now find and block millions of fake accounts every day and detect millions more, often within minutes after creation. This makes running these large networks as part of successful IO much more difficult and has been an important driver in the IO shift from wholesale to retail operations. One example of the improving efficacy of our automated detection: over the past year, the majority of CIB networks that we disrupted involved fake accounts which our systems had already automatically detected and disabled by the time we began investigating.

That's not to say that adversaries have abandoned fake account creation. Rather, we've seen them work harder to slowly build fake accounts into more in-depth "personas." For these networks however, this move is a double-edged sword. Not only does it take more time and resources to get them up and running, but the cost per asset increases every time a network is found and removed and an adversary has to start over.

4.2 Product innovation and adversarial design

Because we know that we face persistent and often well-resourced adversaries, one thing we rely on to anticipate and defend against new ways to exploit our platforms is adversarial design.

Adversarial design is the process of thinking like a threat actor in circumventing our own defenses so we can make our policies, scaled detection, and products more resilient to manipulation techniques, in addition to finding ways to provide people with more context about what information they see on our platform.

Here are a few notable examples of adversarial design in practice:

Authorization and transparency requirements for political and issue ads

Between 2016 and 2018, we saw foreign interference campaigns run ads in other countries — most famously, the IRA paying in rubles for political ads in the United States. By contrast, we now require people to verify that they are in fact based in the country in which they want to run political or issue ads. As a result of this policy change, between March and Election Day in the US in 2020, we rejected social Issues and politics ad submissions 3.3 million times because they didn't complete the authorization process. This is not to say that the ads we rejected were malicious in their intent, but we believe this authorization process added necessary friction for influence operators trying to reach people with paid political messages.

One of the ways to assess the effectiveness of adversarial design is to watch for changes in the behavior of threat actors following the change. We've seen IO operators try to hire locals to run ads for them, leaving a much broader footprint on the ground which exposes them to detection by us, law enforcement, researchers and journalists. Ahead of the US 2020 election, for example, we removed several Russia-linked operations that hired locals in Ghana and the US including to run ads targeting the US.

In addition, we also built a public ad library where political and issue ads remain for seven years, even if the page that posted them is no longer operational. With this searchable archive, journalists, researchers, and the general public can see who is behind a given political and issue ad on Facebook and Instagram, compare the spend of any advertiser running issue or political ads, and even see those ads' potential reach.

State-media labels

As we mentioned in *Section 1 "Defining IO,"* influence operations come in many forms — overt and covert. Over the last several years, we have seen operations that amplified state-media stories, and we have also seen state-controlled media from countries like Russia and Iran publicize content that originated with influence operations. Finally, we also saw covert influence operations run directly by state-controlled media entities.

Without a clear signal that a particular post is authored by a news Page controlled by a particular government, people may treat that information as if it came from a neutral source. To help them know who's behind the content they interact with and judge it with that context in mind, we now label state-controlled media content and Pages and their ads.

Page transparency tools

We've seen influence operations run deceptive Pages that repeatedly switch names and contexts, or pretend to be managed from one country when they are actually managed from another. Consider, for example, a Page that starts off providing local news or entertainment to then switches to promoting a political candidate just months before an election. To combat this kind of misleading activity, we built Page transparency tools that provide additional context, such as the history of changes made by Page admins and where those admins are located.

These transparency tools have repeatedly enabled external researchers and journalists to uncover deceptive behaviors — from run-of-the-mill spam networks to complex influence operations — and flag them to us for investigation. While not every flag leads to a CIB network, we believe in the value of empowering external research and investigative reporting. In addition, this extra context provides a useful frame of reference to the public so they judge what they see on our platform.

Promoting timely and accurate information

We know that influence operations are at their most virulent in information vacuums. That means that our efforts to find and stop IO are most effective when we combine them with enabling access to vetted, accurate information about major societal moments, like the current global pandemic or elections.

Ahead of the US 2020 election, Facebook launched the Voting Information Center ("VIC") that provided people with an accessible source of reliable voting information, tailored for their local area. In it, we highlighted facts about voting to help inform people, helping to inoculate them against false claims about election results or the validity of various electoral methods (*e.g.* the use of mail-in ballots, counting process, *etc.*). From the day it was launched through the election, 140 million people visited the VIC (over 33 million on Election Day alone) and it helped 4.5 million people register to vote in the US.

Protecting highly targeted individuals

Hack-and-leak — where a bad actor steals sensitive information, sometimes manipulates it, and then strategically releases it to influence public debate — has been one of the threats we were particularly focused on and concerned about ahead of the November elections in the US. We anticipate this particular tactic to remain a risk globally.

We know that the ultimate goal of influence operations using this tactic is to drive a particular narrative using media coverage of hacked information. This is why our priority has been to ensure that the accounts of the likely targets for these hacks — political campaigns — are as secure as possible. To provide enhanced protections to this high-target category, we created Facebook Protect for candidates, their staffers and election officials around the country.

4.3. Partnerships with industry, government and civil society

We know that influence operations are rarely confined to one platform. Many of the takedowns referenced in this paper involved information sharing with our peers at technology companies like Twitter and Google, as well as with security researchers, investigative journalists and law enforcement. These partnerships serve as force multipliers, helping us ensure that when one of us detects an IO threat emerges, each of us can investigate it across many platforms.

Whenever we receive a tip from external parties — be it law enforcement or external researchers or other platforms — we conduct our own internal investigation to determine whether we see sufficient evidence of violating activity on our platform to take action in line with our policies. We’ve seen cases whereby activity by the same actor on different platforms manifests differently, including being non-violating. In each case, we take time to investigate the leads we receive before taking any action.

While each platform only has visibility into activity on its respective platform, external researchers and reporters can help connect the dots across platforms, as well as smaller services that don’t yet have dedicated teams looking for these threats. Together, we play a critical role in helping defenders understand the full scope of IO, including its off-platform activity and even presence in the real world.

For their part, governments can provide valuable early warnings when they see threats directed at domestic companies, civic debate or elections, so that platforms can independently evaluate and act as appropriate. In the final weeks before the US 2020 election, tech companies and the US government agencies tasked with protecting the election met weekly in an effort to tackle IO and ensure the integrity of election information across our various platforms. This collaborative approach paid off, allowing us and others to swiftly disrupt foreign and domestic CIB networks earlier in their operations, before they could build their audiences.

4.4 Building deterrence

In addition to proactively looking for and removing IO, part of our security strategy is to find ways to impose costs on threat actors and deter adversarial behavior. Over the past few years, we've aimed to use public transparency and predictability in our enforcement as a deterrent element against IO.

There is reputational cost in being publicly labeled and taken down for foreign or domestic IO. There is also a business risk in losing your company's infrastructure, particularly if its entire business model is built on providing ready-built accounts and Pages to reach target audiences for deceptive purposes.

However, public takedowns by social media platforms and exposure by civil society researchers and media can at times accomplish only so much against determined threat actors. We believe there is a need for a *regulatory* deterrent so we can collectively impose a higher cost on people behind these operations.[27] We've already seen precedents when regulators were able to apply penalties on entities engaged in deceptive behavior in the United States.

While we know that influence operations are unlikely to disappear anytime soon, we also know that in addition to public exposure, legal and financial sanctions can be a powerful tool in our collective toolkit against deceptive campaigns.

[27] Based on studying influence operations over the past four years, our team has outlined recommendations for regulatory and legislative principles against deceptive campaigns: https://about.fb.com/news/2020/10/recommended-principles-for-regulation-or-legislation-to-combat-influence-operations/ Also, see Nick Clegg, Vice President of Global Affairs at Facebook, "Op-ed: Facebook's Nick Clegg calls for bipartisan approach to break the deadlock on internet regulation," CNBC, May 24, 2021 https://www.cnbc.com/2021/05/24/facebooks-nick-clegg-a-bipartisan-approach-to-break-the-deadlock-on-internet-regulation.html [Last accessed: May 24, 2021]

SECTION 5

Conclusion

Since 2017, platforms, researchers, investigative reporters, and governments have made significant progress in detecting and stopping influence operations earlier in their life cycle. But we know that persistent and creative adversaries will continue trying as long as there is a public debate to be influenced.

Here is what we expect to see in the coming months and years:

IO moving into grayer spaces

The scaled techniques we saw in 2016 are now harder to pull off, more expensive, and less likely to succeed. As threat actors evade enforcement by co-opting witting and unwitting people to blur the lines between authentic domestic discourse and manipulation, it will get harder to discern what is and isn't part of a deceptive influence campaign. Going forward, as more domestic campaigns push the boundaries of enforcement across platforms, we should expect policy calls to get harder.

Increased actor diversity

While state-run operations will continue to persist, it's important to remain open and flexible in our detection and response efforts because we know that not all IO is state-sponsored. More domestic, non-state, commercial actors are using the same tactics to influence public debate in their strategic interests. This will lead to more challenging attribution, with more layers of obfuscation between the operators and the ultimate benefactor.

Additionally, the spectrum of inauthentic behaviors is wide. They aren't always focused on politically-motivated campaigns. Many engage in lower-sophistication, higher volume, financially-motivated campaigns, like the clickbait content farms from Macedonia that leveraged real people's content about protests in the US to sell merchandise or drive people to ad farms.

It's particularly important to keep this in mind because if we see state-backed operations behind *every* inauthenticity signal it will inevitably play into the hands of sophisticated threat actors seeking to erode trust in democratic institutions and create a perception of widespread influence operations.

"Weaponization" of uncertainty

We expect to see IO actors continue to attempt to weaponize moments of uncertainty, elevate conflicting voices and drive division around the world, including around major crises like the COVID-19 pandemic, critical elections, and civic protests. Historically though, when we investigate IO targeting these defining moments, authentic voices typically outweigh inauthentic attempts to manipulate public debate.

Because influence operations are rarely limited to one medium, the defender community should include traditional media, tech platforms, democratic government and international organizations, and civil society voices. We expect this to be most difficult in the absence of strong democratic norms around what constitutes authentic discourse.

Here is how the defender community can organize to tackle IO:

- Apply adversarial design across public spaces;
- Develop societal and regulatory norms against influence operations and deception, including by authentic influential voices;
- Distinguish sophisticated influence operations from low-level clickbait, financially-motivated inauthentic behavior, and overt influence activities to calibrate response, and build appropriate enforcement strategies;
- Build and strengthen partnerships to tackle IO comprehensively.

We know that our efforts described in this paper – continuing investigations, automated detection, adversarial product design, partnerships and deterrence – will require constant maintenance and refinement to remain effective. Our adversaries will keep trying to find a way around them. And while we know that their new tactics may be harder to detect for a moment, they are also much harder, more expensive and riskier for them in the long run. Our holistic counter-IO approach and additional transparency has imposed significant friction on the operators, led to over 150 takedowns in the past four years, and created a more hostile environment for influence operations.

To continue imposing friction on IO worldwide, it is worthwhile to deploy adversarial design thinking across all the mediums leveraged by these campaigns (not just tech platforms). That way, over time, by fundamentally changing the nature of the environments they target, we can make deception more difficult, more costly, and higher risk.

There is already a good deal of public discussion about how traditional media and political campaigns can be used by IO to amplify their deceptive narratives, including by way of hack-and-leak operations.[28] Going forward, in order to make these campaigns less effective, we will have to establish new norms and playbooks for responding to IO within a given medium and across society. By shaping the terrain of IO conflict, we can become more resilient to manipulation and stay ahead of the attackers.

Looking back on the US 2020 election, the defenders — including independent researchers, the major social media platforms and the government agencies tasked with protecting the election — were able to detect and disrupt major attempts at IO before they had a significant impact. While this is cause for optimism, one thing remains clear - adversarial threat actors are clever and will stay nimble. And so it will be on all of us in the defender community to adopt the current best practices and formulate some new ones; so that we can stay ahead of the trends we know of now, and the ones which we're yet to uncover in the future.

[28] Janine Zacharia and Andrew Grotto, "How to Report Responsibly on Hacks and Disinformation," Stanford University, 2020 https://fsi.stanford.edu/publication/full-report-how-report-responsibly-hacks-disinformation [Last accessed: May 11, 2021]; Whitney Phillips, "The Oxygen of Amplification," Data and Society Research Institute, 2018 https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf [Last accessed: May 11, 2021]

APPENDIX

List of CIB Disruptions, 2017-2021

This report draws on over 100 CIB networks that we found and disrupted on our platform between 2017 and 2020 with dozens more added since, in 2021. Because over the years, we have shared our findings with relative consistency, this report provided a useful public record of threat evolution and response.

It’s worth noting that this IO Threat Report is limited to known CIB networks. To the best of our knowledge, it is the most comprehensive record of both foreign and domestic operations, including state and non-state campaigns, and therefore provides a useful window into the global nature and trends of IO on our platform. These networks came from over 50 countries and operated in dozens of languages. We continue to grow our global capacity and will keep reporting our findings across various facets of influence operations.

To enable easy access to this record for further research among the defender community, we’re sharing a comprehensive table that includes each CIB takedown we have reported since September 2017. These CIB operations are organized by the date of the disclosure, and the *.csv* file includes links to the original source, the geographic origin of each network, its targets, and the number of assets removed with each disruption at the time of announcement.

Please note that for several earlier operations, precise figures were not available at the time of public reporting, and a few takedowns were reported through the press without a standalone Newsroom post. Over time, we have refined and systematized our reporting to include a consistent set of statistics and descriptions of violating behaviors we found. The table also provides the number of networks we removed as part of each disruption. In a handful of instances, we disrupted multiple networks and announced them in a single composite report.

See the *.csv* file here.

Authors

Nathaniel Gleicher, Head of Security Policy
Margarita Franklin, Head of Security Communications
David Agranovich, Director, Threat Disruption
Ben Nimmo, Global IO Threat Intelligence Lead
Olga Belogolova, IO Product Policy Manager
Mike Torrey, Threat Intelligence Analyst