ISD | Powering solutions to extremism and polarisation

Hoodwinked: Coordinated Inauthentic Behaviour on Facebook

Chloe Colliver
Jennie King
Eisha Maharasingam-Shah

About This Report

This briefing provides an overview of 'coordinated inauthentic behaviour' (CIB) on Facebook. It reviews the information made public on CIB through Facebook's own reporting, assessing the scale of CIB across Facebook and Instagram, the profit Facebook has made from it and the intricacies of the networks themselves. Ahead of the US presidential elections, the briefing highlights the persistent threat of large-scale platform manipulation and provides recommendations for social media platforms if they are to combat CIB on their platforms and protect their users from its detrimental effects.

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org

Executive summary

In September 2020, a recently fired Facebook employee, Sophie Zhang, wrote a memo criticising Facebook for failing to respond effectively to global inauthentic and coordinated political activity on the platform.[1] She raised concerns that researchers and policymakers have highlighted for some time, namely that Facebook enabled political operatives all over the world to conduct deceptive activity targeting elections at enormous scale, with a very low bar for entry.

For two years, Facebook has published transparency reports detailing the actions the platform has taken on what it terms "coordinated inauthentic behaviour" (CIB). Broadly, Facebook's CIB policy outlines activities designated as "inauthentic", in the sense of being covert, deceptive and deliberately misleading, as well as being "coordinated", or involving the use of "multiple Facebook or Instagram assets ... in concert" to engage in inauthentic behaviour.[2] In the wake of continuing evidence of large-scale deceptive networks on the platform over recent months, and the revelations contained in Sophie Zhang's personal account of dealing with these threats for the platform, the Institute for Strategic Dialogue (ISD) conducted the following short review of the information that is publicly available about Facebook's actions to date concerning CIB.

The data tells only a partial story. What information there is shows the residual and significant scale of deceptive activity targeting electorates around the globe on Facebook, from nation states, public relations companies and ideologically motivated hate groups, among others. What it does not and cannot tell us is the true scale of this kind of activity, including that which is not detected or reported by the company. Independent researchers, including from ISD, continue to identify examples of large-scale CIB on the platform, despite having minimal access to Facebook data. The evidence suggests that the examples provided by Facebook over the past two years are merely the tip of a very large and potentially very dangerous iceberg. ISD has provided a set of recommendations at the end of this briefing that suggest ways to understand and respond effectively to the threat of CIB on Facebook, in the hope of being able to deter, detect and deactivate examples of fraudulent and fake activity targeting citizens on the platform around the world.

Key findings of our research include:

- Between July 2018 and July 2020, Facebook removed **78 networks** for CIB, encompassing **23,608 social media assets** (individual pages, groups, accounts) across Facebook and Instagram.[3]

- Between July 2018 and July 2020, Facebook made over **$23 million in advertising revenue** from inauthentic networks that violated the platform's policies.

- The majority of these networks were found via ongoing Facebook investigations, via connections to previously removed networks or from general monitoring activity from the platform's internal detection teams. However, **almost a third of network removals were prompted or helped in some way by an external tip**, either through journalistic reporting, trusted partner flags, intel from a law enforcement agency or alerts from another social media company.

- **The sources most frequently attributed as responsible for CIB networks are Russian and Iranian actors.** It is impossible to know whether this provides any sense of the scale of activity from those actors in comparison with other sources, as it is equally feasible that those states receive greater scrutiny from Facebook and from external researchers or journalists. Unless independent researchers have greater access to Facebook data, CIB removals may remain a case of Facebook "finding what it looks for", in line with its own capacity and priorities, rather than a neutral assessment of what kind of activity exists across Facebook and Instagram.

- The data made available by Facebook suggests that **the US, Ukraine and the UK are the most common targets of CIB network activity**. However, the networks promote narratives that span a large range of geographies, political agendas, wedge issues and audiences.

Introduction

In July 2020, Facebook announced the removal of coordinated inauthentic networks run by allies of the Brazilian President Jair Bolsonaro.[4] The pages, groups and accounts removed had a combined audience reach of over 2 million[5] and were used to promote the Brazilian president, disparage his critics and push misleading and dangerous arguments about the Coronavirus pandemic. Yet such scenarios, whereby large-scale, covert political manipulation is enacted on social media, are now so commonplace that the Bolsonaro case barely made a dent in daily news cycles.

This kind of activity has become commonplace on Facebook, a fact made public in part by the transparency reports about takedowns released since July 2018 under the platform's CIB policy. The policy itself, still somewhat cloaked in mystery, broadly covers activities designated "inauthentic", in the sense of being covert, deceptive and deliberately misleading, as well as being "coordinated". The latter is defined as activity involving the use of "multiple Facebook or Instagram assets ... in concert" to engage in inauthentic behaviour. The stated objective of the policy is to protect users from misrepresentation and to "create a space where people can trust the people and communities they interact with".[6]

The transparency reports are undoubtedly a step forward, providing a top-line overview of policy violations that are both detected and acted on by the company. They offer information on the number of Facebook or Instagram assets (pages, groups and accounts) involved in such networks, as well as the general content themes and, where identified, the actors responsible. But there remain serious gaps and unanswered questions:

- How comprehensive or consistent is Facebook's enforcement of CIB?
- How much transparency is there on decision making and disclosure?
- What data is archived and made available for researchers to understand emerging trends?
- To what extent is Facebook's response to CIB deterring bad actors or preventing re-offence?
- How much money has the platform made (and retained) from these networks, and how is that revenue disbursed?

The frequency of CIB now disclosed means the topic struggles to earn column inches or penetrate public debate; larger cases may be covered, but are quickly eclipsed by other, more sensational crises online. In light of this, ISD has reviewed the overall picture of Facebook's reporting between July 2018 and July 2020. The aim of this briefing note is twofold: to highlight the residual threat of large-scale platform manipulation on Facebook in the final months before the US presidential elections, and to demonstrate the need for greater data access so that researchers can support detection and learn retrospectively from these incidents of platform manipulation.

What is the scale of CIB on Facebook and Instagram?

Between July 2018 and July 2020, Facebook released 34 CIB transparency reports, each detailing examples of networks it has detected itself or been made aware of through independent reporting and research. In total, 78 CIB networks were exposed and taken down in this period, encompassing 23,608 social media assets (individual pages, groups, accounts) across Facebook and Instagram combined.[7]

Despite the steps taken to understand and deal with incidents of CIB, it is fair to assume there is a wealth of activity that slips through the net, evading detection and any form of expert analysis. **Facebook took down over 3.2 billion fake accounts from April to September 2019**, beyond those networks removed under its specific CIB policies.[8] In comparison, there were an estimated 1.6 billion daily active users worldwide in the same period.[9] The overall scale of misleading, false and deceptive activity occurring on the platform is startling. Researchers with almost no access to data on the platform are still finding coordinated inauthentic networks across the globe. All these indicators point to a likely flourishing of CIB on Facebook, of which the platform's reported removals are probably barely scratching the surface.[10]

Does Facebook profit from CIB?

We can begin to gauge the profit made by Facebook from known CIB activity on its platform by manually aggregating the data contained in individual transparency reports, which disclose how much each CIB network spent on advertising and the time period and associated currency (see Table 1). The reports provided by Facebook over the past two years show that **the total revenue for Facebook stood at nearly $23 million by July 2020**, which includes only the revenue directly generated from adverts paid for by the inauthentic networks themselves. While the platform openly reports this information, there is no stated policy on how Facebook manages the revenue once a network is detected and removed.

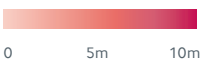
Of the 78 networks removed between July 2018 and July 2020, only 12 did not appear to have associated ad spend, while some networks had generated as much as $9.5 million before their removal. If there is an institutional stance on such revenue, it is not widely known or made available in the public domain.

As Facebook expands its capacity and interest in CIB, it is critical to establish due process (for example, reinvesting the profit in fact-checking efforts or digital literacy programmes), and to disclose such allocations via the company's regular transparency reporting. Regulators, researchers and media outlets should be vigilant in holding platforms to account for this "negative profit".

The largest network removed contained 1,995 platform assets (pages, accounts or groups) across Facebook and Instagram and **originated in Indonesia**, while **the smallest network contained 11 assets** and **originated in Iran**. The relative scale of these networks does not necessarily correlate with revenue for the platform – even **small-scale efforts at online manipulation can generate large ad spends for Facebook, while expansive efforts can have smaller profit margins**. **The Indonesian case mentioned above had a mere $4 in associated ad spend**, while a network of just 23 assets generated ~$1.16 million; the most profitable network had around the median number of assets (927) but had spent over $9.5 million on Facebook ads before its removal.

Table 1 Facebook's advertising revenue compared to size of CIB network between July 2018 and July 2020, by number of assets

Pages, accounts, groups, events removed from Facebook and Instagram	Advertising revenue generated by the network ($)
11	Not disclosed
14	2
15	800
15	70,000
15	Not disclosed
21	Not disclosed
22	>18,000
22	20
23	Around 1.16m
28	10,000
30	>100
31	650
31	>1
32	11,000
35	114,000
44	160
51	Not disclosed
58	Not disclosed
69	11
82	>100
85	Around 1,500
88	1,500
93	About 29,000
97	1
97	Not disclosed
100	1,275
100	77,000
103	1,100
105	About 1.91m
108	>308,000
108	Less than 308,000
114	Not disclosed
119	>400
120	1.93m
122	Not disclosed
122	>450
122	Around 700
126	Not disclosed
137	1,500
140	3,150
145	300,000
148	25,000
153	23,800
169	About 1.38m
174	1.38m
176	>190
200	59,000
203	379
206	>1,600
212	5,800
225	>1,200
226	>20,000
240	930
251	>31,000
265	812,000
362	About 7,600
364	135,000
387	167,000
396	1.6m
397	108,000
418	316,000
432	270,000
448	>150,000
513	15,000
540	1,600
559	Around 216,000
652	<12,000
687	39,000
783	>30,000
790	30,000
810	Not disclosed
927	>9.5m
993	331,000
1,669	<23,000
1,731	>48,500
1,907	Not disclosed
1,995	4



How were these networks identified?

Facebook reporting provides some details on how each network was detected. In examples like the statement by Facebook shown in Figure 1, attribution is clear (a tip from cybersecurity firm, FireEye), but in many cases the framing makes it difficult to gauge the primacy of internal versus external input. Tips or analysis from sources outside Facebook were involved in over a third of the reported networks, although the extent of their role is not always clear. The remainder were discovered through internal Facebook teams who monitor known and emerging threats.

Figure 1 Snapshot of information provided in Facebook CIB reports

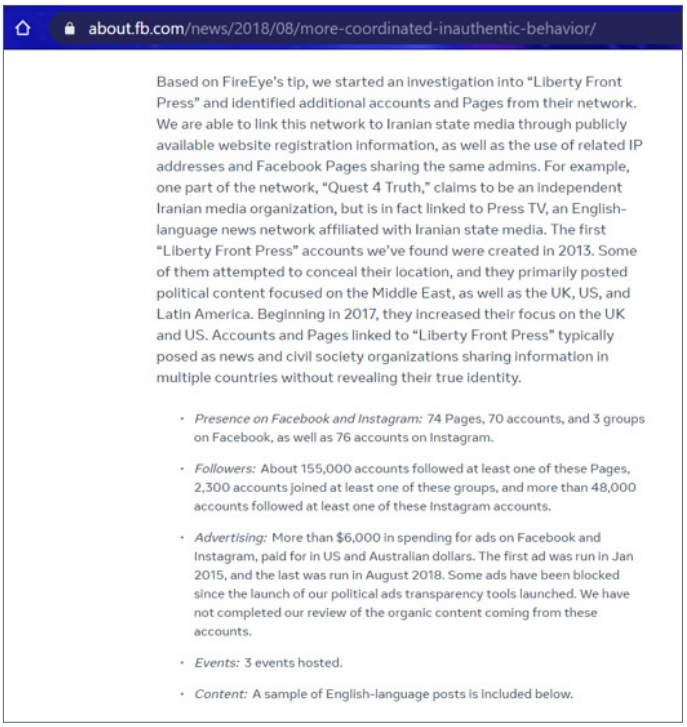


Table 2 shows that **the majority of detections of CIB came via an ongoing Facebook investigation or links to previously removed assets** (n = 25), alongside the general activity of their internal monitoring teams (n = 24). Another **ten networks were associated with the platform's election monitoring** in various contexts globally: one each relating to Indonesia and Ukraine, three in connection to India and the remaining five related to the upcoming US presidential race. An external tip prompted or helped contribute to **29 network removals**, either through open reporting (e.g. published articles that Facebook then actioned); intel shared from a third party (e.g. civil society groups or research bodies); intel from a law enforcement agency (UK and US); or alerts from another social media company (only Twitter cited directly).

Table 2 Motives cited for investigation of CIB on Facebook

Primary explanation provided by Facebook	Incidents (total networks n=78)
Identified by Facebook internal team	24
Identified by Facebook internal team as part of election monitoring	10
Identified by Facebook internal team as part of an ongoing investigation, or linked to previously removed assets	25
Prompted by public media or open source reporting	12
Explicit tip or shared intel from a third party (e.g. research organisations, civil society groups, fact-checkers, journalists, cybersecurity firms)	12
Explicit tip or shared intel from other social media platform	3
Explicit tip or shared intel from law enforcement agency	2
Not specified	3

Note: some investigations cited multiple sources (e.g. a mixture of third-party research and internal monitoring, or an ongoing investigation with insight from public reporting). Thus the overall total in the right-hand column exceeds 78.

What do we know about the networks?

After discovery, data is shared with a small handful of research organisations that partner with Facebook before removals are enacted. These bodies provide an additional review of the pages, groups or accounts in question and include the Atlantic Council's Digital Forensic Research Lab and network analysis firm Graphika.[11] The public can access additional details about the content produced or shared by these networks through this kind of reporting, beyond the characteristics touched on by Facebook CIB reports.

Crucially, no live or retrospective data is shared with anyone beyond those partner institutions for post-hoc analysis. This limits the ability of researchers to tackle CIB in two key ways. First, researchers are unable to study the strategies employed by bad actors, and therefore predict future trends or nuanced detection methods on Facebook or Instagram. Secondly, they cannot explore how such activity maps across platforms (a known component of many CIB campaigns), which could in turn support better cross-platform efforts.

In comparison, Twitter provides datasets of removed networks to vetted researchers, helping deepen understanding of platform manipulation and develop a systemic response.[12] Twitter's archive allows experts to monitor a rapidly evolving set of tactics and narratives, enhanced by their expertise in a given geography, ideology or language. There are some valid concerns about how openly such data can or should be publicly disclosed, considering the risk of copycat behaviour, violations of data privacy or more sophisticated forms of evasion. That said, plenty of models for tiered, anonymised access exist within the tech sector and would provide Facebook and the wider public with a wealth of additional insight to combat CIB long term.

Where did the networks originate?

It can be near impossible for researchers to attribute an inauthentic network to an actor or set of actors when relying solely on open source data. Many networks have complex structures within and across platforms, the costs of shielding identity online are low and the types of data that are most useful to determine attribution on social media platforms are not usually publicly available. Facebook's transparency reports shed some light on the issue of who is behind these types of networks, attributing each to a geographic source, and in some cases a government, company or individual within that context. Staff working for the platform can examine a wealth of data points that are unavailable to external researchers, granting them much greater resources with which to ascertain the "source actors". However, as the then chief security officer of Facebook Alex Stamos wrote in July 2018:

> *The first challenge is figuring out the type of entity to which we are attributing responsibility. This is harder than it might sound. It is standard for both traditional security attacks and information operations to be conducted using commercial infrastructure or computers belonging to innocent people that have been compromised. As a result, simple techniques like blaming the owner of an IP [internet protocol] address that was used to register a malicious account usually aren't sufficient to accurately determine who's responsible.*
>
> *Instead, we try to:*
>
> - *Link suspicious activity to the individual or group with primary operational responsibility for the malicious action. We can then potentially associate multiple campaigns to one set of actors, study how they abuse our systems, and take appropriate countermeasures.*
> - *Tie a specific actor to a real-world sponsor. This could include a political organization, a nation-state, or a non-political entity.*[13]

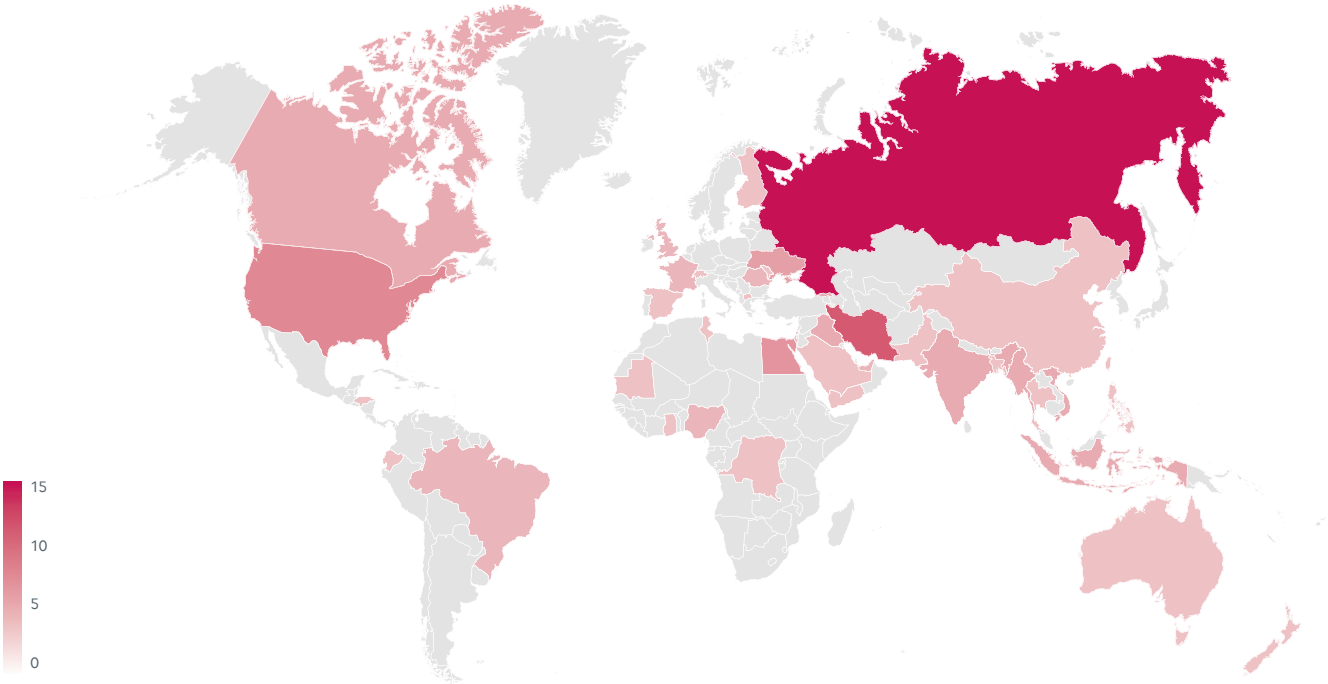
To help determine attribution, Facebook refers to four general categories:

- **Political motivations (inferring links by the known political goals of a given nation state)**
- **Coordination (signs rather than coincidence of proactive networks)**
- **Tools, techniques and procedures (patterns in the methods used)**
- **Technical forensics (studying "indicators of compromise" to establish shared software or infrastructure).**

Sometimes there is evidence in each category, other times the data is too scant or fragmented to draw a conclusion. In many cases, Facebook refrains from making a public claim around who is responsible, at least not naming a specific group.

While we have outlined how many networks were exposed from different geographic origins (see Table 3), such information should be treated with extreme caution when drawing any inferences about the potential scale of activity stemming from different actors. **Most networks were linked to Russia**

Table 3 Stated country of origin of CIB networks using Facebook between July 2018 and July 2020



and Iran (13 and 10, respectively) and while this demonstrates clear and consistent threats originating from those countries, it is equally feasible that those states have garnered a certain reputation and therefore receive greater scrutiny from platforms. Much likely depends on where internal and external researchers have expertise (especially linguistic and geo-political), where previous networks have left clues about future activity or where external pressure from policymakers and the media focuses the platform's attention.

Thus, **CIB takedowns from Facebook to date may be partly self-selecting or hampered by "finding what you look for", rather than a neutral assessment of what exists across Facebook and Instagram**. Without access to underlying trend data, we cannot establish whether detection models are biased towards a particular region or issue set. Indeed, **11 of the 78 networks were linked to multiple countries**, sometimes in regional pairs (e.g. United Arab Emirates and Egypt; Russia and Ukraine; Myanmar and Vietnam), others in more unexpected combinations (e.g. Vietnam and the US; Russia, Ghana and Nigeria; Canada and Ecuador) or broadly global (e.g. one network linked to 12 countries spanning Australasia, Europe, North America and South East Asia).

Stated country of origin	Networks (total n=78)
Russia	15
Iran	10
US	6
Egypt	5
Ukraine	4
Canada, Georgia, India, Indonesia, Iraq, Myanmar, Vietnam	3
Brazil, France, Nigeria, Romania, UK, United Arab Emirates	2
Australia, Bangladesh, China, Democratic Republic of Congo, Ecuador, Finland, Ghana, Honduras, Hong Kong, Israel, Kosovo, Macedonia, Mauritania, Moldova, New Zealand, Pakistan, Philippines, Saudi Arabia, Spain, Switzerland, Taiwan, Thailand, Tunisia, Yemen	1
Multiple countries linked to an individual network	11

Where is CIB targeted?

The countries targeted with the most CIB networks between July 2018 and July 2020 were the US, Ukraine and the UK, with 15, 8 and 5 CIB networks respectively. The networks disclosed by Facebook over these years covered a range of geographies, political agendas, wedge issues and audiences, including content that is pro-, anti- or relevant to:

- Al-Jazeera
- Amnesty International
- astrology
- the Bangladeshi government
- Black Lives Matter
- Brexit
- China
- celebrities and beauty tips
- civil rights movements
- the conflicts in Syria and Yemen (incl. the respective involvement of Iran, Qatar, Turkey, Saudi Arabia and the UAE)
- COVID-19
- Egypt
- election integrity
- elections in Georgia, India, Madagascar, Mozambique and the US
- ethnic divides
- Hezbollah
- the Indian Bharatiya Janata Party and Indian National Congress
- Iran
- Iraq (incl. the US conflict)
- ISIS
- Israel/Palestine
- Jeremy Corbyn (former leader of the UK Labour Party)
- Kurdish politics
- Latin American politics (Argentina, Brazil, Chile, Ecuador, El Salvador, Peru, Venezuela)
- left-wing politics
- LGBTQ+ issues
- Libyan politics
- Muslims in Russia
- Myanmar
- NATO
- political misconduct
- Premier League football
- President Assad
- President Trump
- pro-democracy protests in Hong Kong
- religious beliefs
- Rohingya Muslims
- Russia
- Saudi Crown Prince Mohammed Bin Salman and his political agenda
- The Spanish Partido Popular party
- Sudan (incl. Sudanese-Russian relations)
- Thai politics and activism
- Ukraine (incl. the ongoing situation in Crimea)
- US immigration policy
- West Papua independence movements.

Some networks appeared to stoke division on both sides of a political issue or context, including the use of hate speech and incendiary comments.

It is vital to note that while Facebook may provide sample screenshots of the content removed, **CIB reports do not analyse the narratives spread by a network, or seek to explain its intended goal** (political, social, cultural etc.). The emphasis is on the number of assets, followers, ad spend, associated events and some general context of origin or audience. Reports may also highlight a key tactic (e.g. "owners typically represented themselves as locals… and posted news stories on current events") and overall themes (e.g. "politically charged topics such as race relations"), but do not expand on the stances taken or calls to action. Additional reporting from teams like Graphika and DFR Lab does provide this type of insight, and can help to shed light on the strategy or tactics employed by an operation. Still, without retrospective access to the relevant data, we can only gain an anecdotal snapshot of these networks' end-goals, or how they intersected with wider events at the local or global level.

Implications: What do these reports tell us?

1. Current punitive action does not seem sufficient to break down resilient networks

As outlined above, the largest subset of detections stemmed from known CIB networks and actors, including previous investigations conducted by Facebook. This may suggest that bad actors tend to reoffend, and in turn that content removal alone is an insufficient deterrent for CIB. The barrier to entry may be higher if networks are periodically removed, yet Facebook still fails to tackle the root cause: a group of individuals committed to manipulating platforms and evading detection in ever more creative ways.

In 2020, platforms face an escalating arms race with CIB actors, and could become embroiled in a detection–removal–re-offence cycle with a handful of determined groups. This not only undermines the disincentives imposed by Facebook, but also risks occupying a disproportionate amount of their internal resource and attention.

Facebook has acknowledged this tension, citing an ongoing trade-off between the speed of response and improving platform defences. Chad Greene, Director of Security at Facebook, said in a 2018 article:

> *As soon as a cyber threat is discovered, security teams face a difficult decision: when to take action. Do we immediately shut down a campaign in order to prevent harm? Or do we spend time investigating the extent of the attack and who's behind it so we can prevent them from doing bad things again in the future?*
>
> *Cyber threats don't happen in a vacuum. Nor should investigations. Really understanding the nature of a threat requires understanding how the actors communicate, how they acquire things like hosting and domain registration, and how the threat manifests across other services.*
>
> *This is particularly true for highly sophisticated actors who are adept at covering their tracks. We want to understand their tactics and respond in a way that keeps them off Facebook for good. Amateur actors, on the other hand, can be taken down quickly with relative confidence that we'd be able to find them if they crop up elsewhere — even with limited information on who they are or how they operate.*[14]

Later in the same report, Greene states that action is often taken before an investigation is exhausted, especially if there is an immediate risk to safety. Such a threshold applies not solely to physical harm, but also to "how a threat might impact upcoming world events". Decisions taken may include selecting the timing of punitive action so that a CIB network has minimal time to regroup before a major event (e.g. an election).

Taking punitive decisions against CIB can be complex and they can have various unintended consequences. There is little to no transparency over who makes these difficult decisions, how and why. Furthermore, it is unclear how many removed networks reappear in some form, for example when actors are undeterred by takedown and soon turn to the next available point of entry (e.g. by re-forming private groups, pages and events under slightly revised names; adapting network language or dog-whistles to evade detection by artificial intelligence; purchasing aged accounts or hacking those with weak security protections; and exploiting loopholes in advertising services to amplify their message). In reality, the platform itself may be unaware of every repeat offence – Facebook reports only include the networks staff have identified and removed, but we have no sense of how many continue to slip through the net beyond anecdotal reports and ad hoc research. We cannot gauge the scale of the problem, or make an informed assessment on whether Facebook's response is proportionate or effective without greater access to data for external researchers to detect and report on more of these incidents themselves.

2. Collaboration with external experts would increase the scope, specificity and effectiveness of counter-CIB efforts

The difficulty of breaking down resilient networks as discussed above relates directly to another problem: Facebook is not leveraging the vast knowledge, capacity and commitment of external partners. A significant proportion of takedowns between July 2018 and July 2020 arose as a result of something outside the platform, be it ongoing civic monitoring, a direct tip, public reporting or intel from law enforcement, despite Facebook imposing increasing constraints to the data access it provides to researchers or the public. Thus, CIB could be observed and pinpointed using only the barest minimum data made available to researchers and experts.

It seems undeniable that all stages of Facebook's response to CIB, from system design to detection to analysis, would be enhanced by the platform formalising its links with external bodies. Facebook can still exercise a "trusted third party" rule, setting criteria around which bodies qualify for access and to what extent, but must expand its pool to plug a number of current gaps. Not least, this includes ensuring the breadth of the linguistic capability of its internal teams, and the local nuance around geopolitics or the evolving ideology of extremist factions. Facebook benefits from the work of think tanks, non-governmental organisations, civil rights organisations, investigative reporters and academics, many of which operate with scant resources or project-limited funding. By providing a structured flow of information to relevant groups, the platform could drastically reduce the financial and human strain of its efforts, while reaping the benefits of enhanced investigations.

Greene observes:

> *Academic researchers are also invaluable partners. This is because third-party experts, both individuals and organizations, often have a unique perspective and additional information that can help us. They also play an important role when it comes to raising the public's awareness about these problems and how people can better protect themselves.*[15]

This statement is truer now than ever, as the range of actors engaged in platform manipulation has diversified and globalised. The Kremlin's infamous Internet Research Agency may still have a stake in activity on Facebook, but the playing field for CIB has expanded dramatically since reports first made headlines in 2016. The platform should capitalise on a research community that has the ability to increase the scale and coverage of this kind of detection work but is currently kept at arm's length. This cohort includes researchers who have developed career specialisms in specific dialects, ideologies, political movements, wedge issues, extremist groups or geo-political battles; their expertise should be harnessed and made central to tackling CIB at scale.

3. There remains a gap in cross-platform partnership on CIB and other platform manipulation

In 2018, Greene assured readers that "to help gather [CIB] information, we often share intelligence with other companies once we have a basic grasp of what's happening. This also lets them better protect their own users".[16] This may well be the case, but it is not evident in Facebook's subsequent CIB reporting. Table 1 shows that only three networks were associated with flags by a fellow social media platform, all of which originated from Twitter. This suggests that there is room for increased collaboration on information threats between platforms.

Policy divergence between companies can prevent such efforts from achieving their full potential, as witnessed in the development of the Global Internet Forum to Counter-Terrorism, but this should not deter future efforts. The cross-platform nature of disinformation, CIB and other harms has long been evidenced[17] and suggests the need for more formal partnerships. This could include crisis response protocols (e.g. before elections), or a hash-sharing database for known debunked content and manipulated videos and images.[18] Moreover, cross-sector support may enhance the ability of emerging, smaller and less well-resourced platforms to confront CIB on their sites, building on the systems and expertise of larger companies.

Endnotes

01 C. Silverman, R. Mac and P. Dixit, '"Blood on my hands": a whistleblower says Facebook ignored global political manipulation', Buzzfeed News, 14 September 2020, https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo

02 Facebook, 'Inauthentic behaviour', in Community Standards, 2020, https://m.facebook.com/communitystandards/inauthentic_behavior/

03 Facebook stated in October 2019 that it 'announced and took down' over 50 networks in 'the past year'. Yet in the company's transparency reports from October 2018 through October 2019 (inclusive), the total number of networks removed and reported on was 42. It is unclear why these discrepancies in reporting have occurred. https://about.fb.com/?s=coordinated+inauthentic+behavior

04 N. Gleicher, 'Removing coordinated inauthentic behaviour', Facebook, 8 July 2020, https://about.fb.com/news/2020/07/removing-political-coordinated-inauthentic-behavior/

05 @DFRLab, 'Facebook removes inauthentic network linked to Bolsonaro allies', Medium, 8 July 2020, https://medium.com/dfrlab/facebook-removes-inauthentic-network-linked-to-bolsonaro-allies-5927b0ae750d

06 Facebook, 'Inauthentic behaviour', in Community Standards, 2020, https://m.facebook.com/communitystandards/inauthentic_behavior/

07 Facebook stated in October 2019 that it 'announced and took down' over 50 networks in 'the past year'. Yet in the company's transparency reports from October 2018 through October 2019 (inclusive), the total number of networks removed and reported on was 42. It is unclear why these discrepancies in reporting have occurred. All of those reports can be found on Facebook's website: https://about.fb.com/?s=coordinated+inauthentic+behavior

08 'Facebook removes 3.2 billion fake accounts, millions of child abuse posts', Reuters Business News, 13 November 2019, https://www.reuters.com/article/us-facebook-enforcement/facebook-removes-32-billion-fake-accounts-millions-of-child-abuse-posts-idUSKBN1XN2B2

09 J. Clement, 'Number of daily active Facebook users worldwide as of 2nd quarter 2020', Statista, 10 August 2020, https://www.statista.com/statistics/346167/facebook-global-dau/

10 The claims made in S. Zhang's memo, reported in Buzzfeed in September 2020, support the argument that there likely remains a significant amount of CIB on Facebook that goes unreported and without timely response from the platform. See Silverman et al., '"Blood on my hands"'.

11 @DFRLab, 'Why we're partnering with Facebook on election integrity', Medium, 17 May 2018, https://medium.com/dfrlab/why-were-partnering-with-facebook-on-election-integrity-19f0ca39db2e

12 Twitter Safety, 'Disclosing new data to our archive of information operations', Twitter, 20 September 2019, https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html

13 Alex Stamos, 'How much can companies know about who's behind cyber threats?', Facebook, 31 July 2018, https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/

14 Chad Greene, 'When to take action against cyber threats', 21 August 2018, Facebook, https://about.fb.com/news/2018/08/more-coordinated-inauthentic-behavior/

15 Ibid.

16 Ibid.

17 For example, see B. Nimmo, C. Francois, L. Ronzaud and C. S. Eib, 'Spamouflage dragon', Graphika, April 2020, https://graphika.com/reports/return-of-the-spamouflage-dragon-1/

18 Joint Tech Innovation, 'Hash sharing consortium', Global Internet Forum to Counter Terrorism, [2020], https://www.gifct.org/joint-tech-innovation/

ISD | Powering solutions to extremism and polarisation

Beirut | Berlin | London | Paris | Washington DC

www.isdglobal.org