The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations

Camille François and evelyn douek

Abstract. Intense public and regulatory pressure following revelations of Russian interference in the US 2016 election led social media platforms to develop new policies to demonstrate how they had addressed the troll-shaped blind spot in their content moderation practices. This moment also gave rise to new transparency regimes that endure to this day and have unique characteristics, notably: the release of regular public reports of enforcement measures; the provision of underlying data to external stakeholders and, sometimes, the public; and collaboration across industry and with government. Despite these positive features, platform policies and transparency regimes related to information operations remain poorly understood. Underappreciated ambiguities and inconsistencies in platforms' work in this area create perverse incentives for enforcement and distort public understanding of information operations. Highlighting these weaknesses is important as platforms expand content moderation practices in ways that build on the methods used in this domain. As platforms expand these practices, they are not continuing to invest in their transparency regimes, and the early promise and momentum behind the creation of these pockets of transparency are being lost as public and regulatory focus turns to other areas of content moderation.

1 Introduction

"Content moderation" is an umbrella term for the way platforms write and enforce their rules for what people can do and say on their services. [1] Despite what the term might suggest, a significant amount of content moderation that platforms perform has little to do with the *content* of posts, videos, images, or messages, but focuses instead on the *actors* behind them or their *behavior* in determining whether something should or should

1. Content moderation is "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann 2015).

doi:10.54501/jots.v1i1.17

not be allowed to remain on their sites.[2] Behavior-based content moderation is as old as web-based services, and has its roots in spam detection and enforcement. Similarly, actor-based moderation has been a key tool in platforms' ability to tackle the spread of violent extremism and other violent organized actors on their services, especially since platforms began to address how ISIS exploited their services (Berger 2015; Bickert and Fishman 2017).

These kinds of behavioral- and actor-based content moderation categories have rapidly multiplied recently, with platforms rolling out more policies to address a broader variety of ways their services are abused by coordinated groups of actors in ways that will not be apparent at the level of individual posts, videos, images, and so forth. This is in no small part a response to the twin high-pressure information events of the COVID-19 pandemic and the US 2020 election, which put platforms and their content moderation practices under the microscope. Platforms scrambled to address issues that became focal points of public and regulatory pressure in 2020, including conspiracy theorists, domestic disinformation networks, and groups using their online services to coordinate offline violence and aggression.[3] The events of 2020 transformed content moderation like so much else in society, with platforms taking a far more interventionist approach to governing their services and addressing the ways they can be used to cause harm online and off.

A previous similar moment of public and regulatory pressure created the most high-profile form of behavioral content moderation. In the wake of the 2016 US election, extraordinary revelations about the extent to which Russia exploited social media platforms to run influence operations spurred a public reckoning in the United States about the role these platforms play in society. Congressional hearings, political ire, and societal pressure ushered in new platform policies directed at dealing with influence operations. It also created novel transparency regimes intended to shed light on what platforms were doing to address the way their services had been exploited. The policies and transparency regimes that emerged from this moment remain unique, and this pocket of information operations-related disclosures is still one of the most transparent in terms of the data provided to the public and outside parties about actions platforms have taken. These regimes also borrow from established practices across the information security context ("infosec," or cybersecurity), not just the content moderation realm.

While there are true benefits from these developments, this article will also show that the way these transparency regimes operate can also create perverse incentives for platforms in how they find and report on these types of abuse and skew public understanding of online information operations more generally (douek 2020). And, importantly, the routine disclosures created by these transparency regimes have not ushered in a new age in platform transparency: platforms have not continued to invest in these regimes to the same degree as pressure decreased with distance from the election, nor have these practices expanded to other key and bordering areas of content moderation. Furthermore, the persistent lack of convergence among platforms on these topics suggests these pockets of transparency may not continue to deepen, nor their scope further clarified.

This article explains these dynamics by tracing the history of the behavioral content moderation practices designed to tackle information operations and the corresponding transparency regimes that platforms created following the public reckoning around the Russian large-scale interference in the 2016 US presidential election. Platform efforts to tackle foreign interference and this specific type of online abuse have come a long way in the intervening years, and the absence of any significant online foreign interference

2. One of us has described this as the "ABC Framework" of content moderation in the information operations space (François 2019).

3. See Section 6 on page 17.

campaigns aimed at disrupting the 2020 presidential election offers a striking contrast to the industry's unpreparedness in 2016.[4] This article, however, does not focus on the question of whether the changes platforms made have been effective in addressing foreign interference or protecting elections from information operations. Instead, our focus is the specific *transparency reporting regimes* for the categories of abuse related to information operations that appeared across the industry in 2017, and how those regimes have held up and evolved.

Accordingly, our focus here is narrow. We do not address transparency practices in content moderation more generally, nor try to answer the fraught substantive question of how to draw the lines between permissible and impermissible online behavior. Our focus is limited to the policies designed to tackle information operations and what platforms disclose to the public about it. Increasingly, the disclosure regimes that exist for information operations takedowns[5] run the risk of becoming isolated accidents of history in response to the extraordinary conditions following the 2016 US election rather than the beginning of a new era of transparency. So far, new categories of behavioral- and actor-based content moderation by platforms have proliferated, but platforms' associated transparency practices have not. This article considers the strengths and weaknesses of the limited regimes created particularly by major platforms to tackle information operations, which carry different names and have different boundaries: "Coordinated Inauthentic Behavior (CIB)" at Facebook, "Information Operations (IO)" at Twitter, and "Coordinated Influence Operation Campaigns (CIOC)" at Google (collectively referred to in this article as CIB/IO/CIOC).[6]

Despite their different coverage,[7] the CIB/IO/CIOC policies share a common origin story and a focus on deceptive activities by organized actors, and are all used to target networks and clusters of accounts rather than single users or single posts. Their corresponding disclosure regimes form a fragile pocket of transparency in the content moderation industry. On closer examination, these disclosure regimes also show the ambiguities and inconsistencies of these policies, and in turn can distort platform incentives and broader public understanding of online disinformation.

We proceed as follows. Section 2 begins by tracing the way that the aftermath of the 2016 election led platforms to provide an unprecedented form of transparency into how their services had been exploited by the IRA and the actions they took in response. Section 3 describes how these disclosures coalesced into formalized transparency regimes that persist to this day. Section 4 shows the persistent and underappreciated ambiguities in each platform's information operations policies and disclosures. Section 5 critically examines the ways these ambiguities can create perverse incentives for platforms and other stakeholders, and blind spots for observers of the online information operations ecosystem. Section 6 shows how these problems are set to become even more acute by cataloging the ways in which platforms are expanding their behavioral content moderation practices without providing the same kind of transparency they do

4. "It's difficult to rigorously compare foreign interference campaigns in the 2016 and 2020 US election cycles, given the enormous differences in awareness and preparedness between both electoral cycles. In 2016, information operations on social media were a true blind spot for entities charged with protecting the integrity of the election" (Election Integrity Partnership 2021, 106–7).

5. "Takedown" is a generic term that has come to describe enforcement action taken against batches of accounts engaged in an operation.

6. We focus on these platforms given they have created the most formal transparency regimes in this area. Our arguments apply more broadly, and in many ways more forcefully, to other platforms that have not created such formalized regimes at all. TikTok, for instance, is briefly mentioned in this article as having adopted Facebook's "Coordinated Inauthentic Behavior" terminology for their own policy tackling information operations, but has done so without adopting the corresponding and regular disclosure policy that Facebook pioneered.

7. See Section 4 on page 11.

for CIB/IO/CIOC.

Our view is that transparency is a normative (if instrumental) good that helps bring accountability to decisions platforms make that have a profound effect on public discourse. But we accept this is not an uncontested view, and it is also not one necessary for our argument in this article.[8] The fact is that platforms *have* developed a distinctive transparency regime for the actions they take against information operations on their services and that the ambiguous and somewhat arbitrary nature of these regimes has underappreciated upstream and downstream effects. This article focuses attention on these effects. Such a focus is important as platforms increase the range of behavioral content moderation they engage in and as lawmakers around the world are writing regulations attempting to make these actions accountable. In making this argument, this article also provides the first detailed account of how these policies have emerged in the wake of the revelations that the Russian Internet Research Agency (IRA) leveraged social media to target the US presidential election. It is especially remarkable that a national security crisis led to a rare pocket of transparency within the industry, as national security considerations typically lead institutions to err on the side of secrecy rather than on that of transparency.

2 The IRA Disclosures: A Pivotal Moment in Content Moderation History

The revelations that the Russian IRA had strategically used Facebook, Google, Twitter, Reddit, Medium, and other platforms to influence the 2016 US presidential election (DNI 2017) kicked off a "techlash," especially in the United States, which in material ways has never abated (Foroohar 2018; Economist 2018). This section gives an account of the events that followed, the public disclosures platforms made into their investigations, and what they found on their services. This history illustrates how ad hoc, unplanned, and contingent the story is.

Public and political outrage in the wake of the events of 2016 led to an unprecedented level of scrutiny on platforms, leading to the first of what has now become a familiar sight of executives from Facebook, Google, and Twitter on Capitol Hill to face bipartisan ire for the way they had allowed their services to be exploited (Kang, Fandos, and Isaac 2017). This questioning and scrutiny brought to light the troll-farm-shaped blind spot in major platforms' content moderation practices that the IRA had taken advantage of (François and Lin 2020). Before the election, platforms generally had no clear policies against these types of operations and no internal teams formally in charge of preventing them.

In April 2017, as lawmakers' concerns kept escalating, Facebook released a white paper titled "Information Operations and Facebook" that addressed what the company had done to tackle this blind spot up until then (Weedon et al. 2017). The white paper, authored by three members of the Facebook security team (Jen Weedon, Alex Stamos, and Will Nuland), laid out how the security team would take ownership of combating this kind of threat going forward, "expand[ing] [their] focus from traditional abusive behavior, such as account hacking, malware, spam and financial scams, to include more subtle and insidious forms of misuse, including attempts to manipulate civic discourse and deceive people" (ibid., 3). The paper included a few screenshots and high-level figures to illustrate how Facebook was beginning to define these forms of misuse and "information operations." The authors also included a shy, roundabout sentence suggesting that Facebook was able to attribute at least part of this activity to Russian entities: "Facebook

8. On the idea that transparency is not always an unmitigated good, see, e.g., Pozen (2020).

is not in a position to make definitive attribution to the actors sponsoring this activity, … however our data does not contradict the attribution provided by the US Director of National Intelligence in the report dated January 6, 2017" (ibid., 11). The DNI report Facebook was referencing had assessed "with high confidence that Russian President Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election" (DNI 2017, 1). It is worth noting that Facebook's position on whether it would publicly attribute these types of campaigns radically changed in the months following this initial report, and the company became much less coy. A little over a year later, in July 2018, the company's chief security officer even published a blog post setting out the contours of Facebook's methodology in attributing information operations (Stamos 2018b).

Pressure on platforms to share what they knew about the extent of Russian influence continued to increase throughout 2017, especially after the Mueller probe into Russian interference in the 2016 US election became a key focus of public debate (Ruiz and Landler 2017). Platforms were building the plane while flying it: methods of fully investigating, defining, and analyzing information operations were new, investigations routinely took months, and new findings from one platform often only came to light as more information surfaced from others' own investigations. What came out of these processes was a slow-drip and scattered set of IRA-related disclosures.

In September 2017, Facebook expanded on its initial white paper, publishing a blog post with further details on the number of accounts and ad spend related to the activity it had uncovered (Stamos 2017). Two weeks later, Twitter followed with its first announcement explicitly addressing Russian interference and disclosing its preliminary findings of how Russia had used its platform to target the US election. This announcement also acknowledged how the platforms were using each other's findings to inform their own investigations and disclosures, stating, "of the roughly 450 accounts that Facebook recently shared as a part of their review, we concluded that 22 had corresponding accounts on Twitter" (Twitter Public Policy 2017).

A month later, on October 30, 2017, (a day ahead of a Congressional hearing on the matter [Kang, Fandos, and Isaac 2017]) Google finally publicly acknowledged its own investigation into IRA activities on its platforms and services (Walker and Salgado 2017) and shared a short summary of its findings. Earlier that month, Pinterest had also acknowledged that it had been affected by the IRA's campaign and had conducted an investigation, finding that, while Russian operatives did not appear to have posted to Pinterest directly, associated content from other platforms like Facebook had ended up on Pinterest through users cross-posting it (Dwoskin 2017). In January 2018, Twitter updated its previous disclosure over two consecutive days, announcing that it had notified approximately 1.4 million people who may have engaged with the IRA accounts (Twitter Public Policy 2018).

In 2017, a few social media platforms (Twitter, Facebook, and Tumblr[9]) briefly adopted the position that users who had directly interacted with these campaigns should receive direct notifications they had done so. That practice was short-lived, and today none of the major social media platforms routinely notify users who have directly interacted with information operations on their services. The original notifications echoed best practices from the cybersecurity field rather than standard practice in content moderation, emulating the state-sponsored attack alerts that Google pioneered in 2012 and that

9. In Congressional hearings, Senator Blumenthal urged major companies affected by the IRA's campaign to adopt these sorts of user-facing disclosure practices. Google did not, and told Sen. Blumenthal in a December 2017 letter that the company would not be able to espouse a similar notification because "content is accessible regardless of whether or not a user is logged in," meaning it "would not be able to identify all those who watched a particular video" (Romm 2017).

others throughout the industry have adopted to warn users whose accounts are being targeted by state-sponsored actors. There are unfortunately no published accounts of the short experiment of bringing this practice to the field of information operations: it is unclear if these notifications were swiftly retired because they are impractical to manage for platforms, confusing for users or otherwise ineffective,[10] or something else altogether.

Over the next few months, the public record of how the platforms had been exploited in the run-up to the 2016 election would slowly become more detailed as the media, researchers, and US governmental entities continued to publicly investigate and expose the scope of the IRA's efforts. On February 16, 2018, the US Department of Justice (DoJ) filed a detailed indictment against the IRA (Kahn 2018). In the accompanying press release, Deputy Attorney General Rod J. Rosenstein noted that the DoJ "received exceptional cooperation from private sector companies like Facebook, Oath, PayPal, and Twitter" (Department of Justice 2018).

This set of disclosures created mounting pressure on other platforms to be transparent about whether their services were also affected. In March 2018, Reddit publicly acknowledged it had been conducting its own investigations, sharing on the subreddit r/announcements its first disclosure of IRA activity. The company acknowledged its lack of transparency and the role of public pressure in prompting its announcement:

> In the past couple of weeks, Reddit has been mentioned as one of the platforms used to promote Russian propaganda. As it's an ongoing investigation, we have been relatively quiet on the topic publicly, which I know can be frustrating. While transparency is important, we also want to be careful to not tip our hand too much while we are investigating. We take the integrity of Reddit extremely seriously, both as the stewards of the site and as Americans. (u/spez 2018)

Tumblr followed shortly afterwards with their own disclosure (Tumblr Staff 2018), announcing that the platform would create a public archive of usernames "that we have determined were engaged in state-sponsored disinformation and propaganda campaigns" (Tumblr Help Center 2018). Although the archive contains less detail than later iterations of such archives, this was the first example of a platform itself providing a public archive of data on information operations.[11] Tumblr's announcement also noted that it had emailed users affected by the Russian campaign, specifically notifying them which fake accounts they had engaged with.

On April 3, 2018, a year after the initial white paper on *Information Operations on Facebook*, Alex Stamos published a Facebook Newsroom post entitled "Authenticity Matters: the IRA has no place on Facebook." (Stamos 2018a) The post details a further takedown of IRA accounts by Facebook that day. The post noted that the action was taken purely because of the *actors* involved, regardless of the type of *content* they shared:

> The IRA has repeatedly used complex networks of inauthentic accounts to deceive and manipulate people who use Facebook, including before, during and after the 2016 US presidential elections. It's why we don't want them on

10. A note from one of us who is passionate about these warnings: The "infosec" or traditional state-sponsored warnings typically contain resources for users to bolster their account security, as stronger security practices help deter these types of threats. It's more complicated to think of calls to actions that would help users better mitigate further threats of being exposed to information operations, making the question of how to measure effectiveness a difficult one.

11. Twitter would launch a more detailed archive in October 2018, which remains the most comprehensive such database, but its claim that this is the "first archive in the industry" is not technically correct; see "Information Operations" (2018).

> Facebook. We removed this latest set of Pages and accounts solely because they were controlled by the IRA—not based on the content. (ibid.)

Later, platforms would come to insist that the key criterion for policing information operations on their platforms was *behavior* (François 2019; douek 2020). Facebook's "coordinated inauthentic behavior" concept (discussed further below) would appear a few months later (Gleicher 2018). But in early 2018, the focus was still on the IRA as an *actor*. Stamos's post was accompanied by an "IRA Takedown Facts" table, sharing a few data points on the number of accounts deactivated that morning and offering the option to download a folder containing "samples" of the activity in the form of three PDFs and two JPEG files with examples of posts and ads the IRA had shared on the platform.

Facebook, Twitter, and Google also provided additional data on IRA-related activity on their platforms to the US Senate Select Committee on Intelligence (SSCI). These disclosures were the basis of two reports by independent researchers documenting the nature and extent of the IRA's 2016 campaign, (Fogel 2018; Howard et al. 2019; DiResta et al. 2019) released by SSCI in December 2018 (SSCI 2018).

Together, these disclosures comprise the public record of the IRA's 2016 campaign. The existence of this corpus may seem fairly unsurprising in hindsight, when similar disclosures are now routine for a number of platforms. At the time, however, these scattered updates represented unparalleled and unprecedented transparency from companies that had historically been extremely opaque about content moderation generally.[12] To this day the information released about the IRA campaign in 2016 remains the most extensive cross-industry investigation and acknowledgment of an influence operation, even if it is made up of a patchwork of unwieldy, uncoordinated, and ad hoc disclosures based on what each company had visibility into and was willing to disclose.

3 From Ad Hoc Disclosures to a Formalized Policy Regime

While the public record of the IRA's 2016 campaign remains the most extensive to date, the pressure on platforms to reassure the public and lawmakers that they were working to prevent similar exploitation of their products going forward led various platforms to create formalized transparency regimes for information operations more generally. This section details this process and the transparency regimes that came out of it.

Generally speaking, the extent of public disclosure about the Russian information operations in 2016 is the exception that proves the rule of opacity in behavioral content moderation. Prior to that, similar campaigns had unfolded in many other countries, but none had resulted in platforms providing the same level of transparency. The 2017 #MacronLeaks campaign provides a particularly stark contrast because it immediately followed the 2017 reckoning with the role that the Russian IRA had played in targeting the US 2016 election, but predated the formalized policy regime that would result from it. It is now well established that an operation originating in Russia targeted the 2017 French election, using strikingly similar methods to those observed in the 2016 US election (Vilmer 2019). But no systematized process of platform public disclosure of the discovery of such operations yet existed, and French lawmakers did not engage in an effort to uncover the scope and nature of the operation comparable to what had occurred in the US. As a result, no detailed information about this operation has ever been published by any of the platforms or made available to researchers. This leaves French researchers at a significant disadvantage as they work to detect potential foreign

12. See, e.g., Singh and Doty (2021) detailing how the major platforms only started releasing any transparency reports about their terms of service enforcement in 2018.

interference efforts targeting the French 2022 presidential election: they don't have access to any public blueprint or concrete examples of the tactics, techniques, and procedures deployed in 2017.

By the end of 2018, Facebook, Twitter, Google, and Reddit would all formalize the ad hoc transparency disclosures and new rules that had been developed with respect to the 2016 Russian influence operation into continuing policy and disclosure regimes for similar types of activity. With different frequency and levels of granularity, each of these major platforms settled on new and separate disclosure practices for this type of activity.

In July 2018, Facebook introduced the concept of "Coordinated Inauthentic Behavior" for the first time as it announced further action being taken against the IRA ("Removing Bad Actors on Facebook" 2018). A few months later, Facebook released an "explanation" of the policy, describing CIB as groups of Pages or people working together to mislead others about who they are or what they are doing (Gleicher 2018). A steady drumbeat of "CIB disclosures" followed, at first on an ad hoc basis, but ultimately, in March 2020, Facebook announced that it would begin publishing a monthly "CIB report" containing details on enforcement actions taken in the prior month under the CIB policy. These reports do not contain a complete record of the different accounts and posts Facebook has taken down, but they do contain top-line figures of the number of "assets" (accounts, Pages, and Groups) and pieces of content Facebook removed as well as example posts. Access to the Pages prior to them being taken down by the platforms is at times granted to third-party research groups, like Graphika or the Digital Forensic Research Lab (DFRLab) at the Atlantic Council, who then document the campaigns and produce public reports and analyses of these takedowns.

In October 2018, Twitter announced it would create an archive for content related to information operations and release "all the accounts and related content associated with potential information operations [found on Twitter] since 2016" (Gadde and Roth 2018). This represented a significant move by Twitter to give researchers access to comprehensive datasets about information operations, enabling a broader array of external research. To this date, Twitter's archive remains the only one of its kind. We discuss below why this approach, while deep, is also narrow because of Twitter's particular definition of the "information operations" it includes. Twitter has since regularly announced additions to the archive when it finds additional operations and shares early access to the data with research groups such as the Stanford Internet Observatory for their investigation and analysis (Twitter Safety 2021). Platforms have not explained how they choose which outside groups to partner with and provide such data to.

In October 2019, Reddit, which had started sharing occasional updates on Russian and Iranian information operations, announced that these announcements would now live in a quarterly security report (u/KeyserSosa 2019). Reddit's Chief Technology Officer acknowledged that the industry trend towards transparency had influenced its own practices: "I would like to acknowledge the reports our peers have published during the past couple of months (or even today). Whenever these reports come out, we always do our own investigation." (ibid.)

It was not until May 2020 that Google announced that it too would move from ad hoc disclosures to a regularly scheduled reporting structure on coordinated influence operations, through the Threat Analysis Group's Quarterly Bulletin (Huntley 2020b). These bulletins contain short, high-level descriptions of enforcement actions taken against "coordinated influence operation campaigns" on Google platforms (including YouTube). In a sentence or two, the bulletins enumerate how many assets were tied to an operation that had been discovered and where the actors behind the operation originated

from. Sometimes the bulletins contain a sentence-long description of the content of the campaign (for example, "[t]his campaign posted content in Arabic about the Syrian civil war and critical of US foreign policy" [Huntley 2020a]), but they never include sample posts or additional data about the accounts involved. They regularly include a reference to disclosures from other platforms (for instance, stating that what Google found was "similar" to the findings reported by another platform [ibid.]) or to entities who have provided leads to Google (for instance, "we received leads from the FBI that supported us in this investigation" [ibid.]).

There are significant disparities in the way each platform reports on their CIB/IO/CIOC enforcement actions, and little explanations for why some platforms offer more detail than others. This could be a product of organizational structure or culture, the choices of individual decision makers inside each company, a part of their broader political or communications strategy, a feature of differences in products affected by these campaigns and resulting data available, or (as is likely) some combination of all of the above. Some platforms may have chosen to attempt to avoid scrutiny by providing less public transparency, while others may hope that greater transparency will engender more trust in their operations. Whatever the reasons, these choices shape the overall field concerned with the study of these campaigns and regulatory and public understanding of online influence operations more generally.

	Facebook	**Twitter**	**Google**	**Reddit**
Terminology	*Coordinated Inauthentic Behavior*	*Information Operations*	*Coordinated Influence Operations Campaigns*	*Suspected Manipulation*
Cadence	Monthly	Ad hoc	Quarterly (but updated on a monthly basis)	Quarterly
What is disclosed?	The report contains short descriptions of the type and number of assets[a] in each campaign, sometimes accompanied by screenshots. Datasets are routinely shared with external researchers who provide longer independent reports to accompany the disclosures. In certain cases, Facebook itself provides additional context on the campaign. Attribution to a specific actor is shared when the company has a high level of confidence in their assessment (Stamos 2018b).	When a new campaign is investigated, the related accounts and posts are added to Twitter's public archive. A hashed version of this content is publicly available through this dedicated site. An unhashed version can be made available to researchers upon request and is routinely shared with research groups ahead of the public announcements. Attribution to a specific actor is shared in the disclosure when Twitter determines it can "reliably make" such an assessment (Twitter Safety 2019).	Very short descriptions of the activity and the number of assets involved, and no additional data provided. Attribution to a specific actor is sometimes shared, although Google hasn't commented publicly on its standards for doing so.	The "Suspected Manipulation" section of the security reports includes a few highlights on the state of policies, tools, priorities, and findings from the team working on these issues. It is less focused on specific campaigns (although it includes a link to disclosures when applicable, usually along with the list of usernames affected by the action and their "karma" distribution). Because the reports are shared in posts on Reddit, the authors often engage with Reddit users in comments, answering questions from the Reddit community and sharing additional details.
Total campaigns disclosed as of 08/31/2021 (source: Disinfodex[b])	157	38	65	3

a. In this context, an "asset" is any type of social media property that is directly tied to this activity: a Page, an ad, a Group, a profile, etc.

b. "Disinfodex is a database of publicly available information about disinformation campaigns. It currently includes disclosures issued by major online platforms and accompanying reports from independent open source investigators"; see www.disinfodex.org.

An unspoken but important expectation created by these regularly scheduled updates is that *if a campaign has been found, it will be disclosed in these reports*. The absence of disclosure therefore suggests a campaign has not been identified by the platform. This assumption is at times erroneous. Facebook's monthly CIB reports make this commitment explicit by prefacing each report with a statement that "as part of our regular CIB reports, we're sharing information about *all* networks we take down over the course of a month" (emphasis added). But there is no way of verifying this and it is unclear that this practice applies for other disclosure regimes throughout the industry: at the time of writing, for instance, the Twitter Information Operations archive has not been updated for about eight months despite the Twitter threat intelligence team continuing to tackle these forms of threats.[13] The opacity of Google's definition of what its CIOC policy covers makes it impossible to tell what should be included in its bulletins and whether they cover all CIOC discovered by Google, or simply a selection of them.

In these disclosures, platforms also routinely acknowledge third parties (including other platforms, investigative reporters, civil society groups, cybersecurity firms, and government partners) as having provided tips or analysis in their investigation—again, a practice more common in the cybersecurity context than in the content moderation one.[14] This provides limited transparency into the sources of leads platforms use to guide their investigations and enforcement. Together, these disclosures show the increasingly routine information sharing between the US government and social media platforms and between platforms themselves. This raises important issues, beyond the scope of this article, about who gets a seat at the decision-making table and why such collaborations involve the same select group of actors again and again.[15]

This is the state of play at the time of writing. Ad hoc disclosures in the aftermath of the 2016 US election crystallized into formal transparency regimes, some of which include data about accounts and posts within these networks, examples, or, at minimum, descriptive narratives of operations found. These regimes represent some of the most comprehensive disclosures across any category of content moderation in the level of detail about what platforms took down, with specific examples; how the relevant assets were discovered; and how that particular network operated. These regimes are the result of a number of stars aligning. The immense public and regulatory pressure was certainly a key driving factor, as was the fact that combating "foreign" interference has long attracted fewer censorship concerns than the idea of policing domestic political speech (douek 2020a). Crucially, the accounts and posts swept up in these categories were, when the category was confined to the original IRA campaign, fake or inauthentic. This alleviated privacy concerns that platforms often (and with some justification) point to as preventing detailed public disclosures about activity on their services.[16] These definitions have then slowly expanded to include authentic accounts when they play a key role in coordinating the inauthentic accounts involved in the operation—it remains true, however, that the bulk of the CIB/IO/CIOC disclosed to date concern inauthentic (fake) accounts.[17]

But behind this apparent boon are a number of factors that obscure the way that this

13. The writing for this article concluded in October 2021, when the latest dataset uploaded to the Twitter archive dated back from February 2021. In this period, Twitter did take action against state-sponsored operations that do not appear in the Information Operation archive, such as the Russia-linked "NAEBC" campaign, which had a small presence on their platform. See Stubbs (2020).
14. Advocates regularly ask for more transparency around the third parties that platforms routinely partner with in other content moderation areas as well: this question often arises in debates on defining hate speech, hate groups, and terrorist organizations, for instance, and which governments, local, or regional partners platforms consult.
15. One of us has called this the rise of content cartels (douek 2020a).
16. See, e.g., Clark (2021).
17. See, e.g., Facebook's removal of Roger Stone's account for ties to CIB (Alba 2020).

regime also distorts public understanding of the information ecosystem as a whole and creates a number of perverse incentives to both over- and under-categorize certain online activity as coordinated, inauthentic behavioral platform manipulation, or as information operations.

4 The Persistent Ambiguities of Existing Definitions

A lot of ink has been spilled on the difficulties of defining disinformation.[18] The policy category that platforms created to deal with information operations following the IRA disclosures was not intended to solve this problem, but to be a narrower more clearly delineated category of online behavior. In theory, because platforms' policies were crafted in response to the same original campaign, there would be a measure of consistency and determinacy across platforms' definitions. In practice, however, as these policies have been developed and applied, platforms have taken different routes, and the ambiguities and inconsistencies of these policies have become apparent. This section details these underappreciated features of how platforms police and disclose information operations on their services.

Each platform operates its own intricate and evolving definition of what falls within the category of problematic platform manipulation. Observing different definitions on different platforms isn't unusual, but what is more uncommon is how little public and regulatory understanding there is about the level of divergence. Facebook coined the policy moniker "coordinated inauthentic behavior" and, in no small part thanks to the platform's communications strategy and investments in publicizing their efforts, the term has become a generic stand-in for the overarching category. Even though Twitter and Google use different monikers in their rules, members of Congress have praised representatives from Facebook, Twitter, and Google for their efforts in tackling "coordinated inauthentic behavior" like it is a technical and objective category (douek 2020c). TikTok, meanwhile, has picked up "CIB" and uses it in its rules without explaining what it considers the term to mean (TikTok 2021). Other platforms have also invoked the term in various statements,[19] but this general and colloquial use of the phrase does not have meaningful content.

As noted above, Facebook's public definition of coordinated inauthentic behavior is "working in concert to engage in inauthentic behavior … where the use of fake accounts is central to the operation" ("Community Standards: 22. Inauthentic Behaviour" 2020). Facebook's attempt to give more detail as to how it applies the policy largely fails to provide clarity, unhelpfully showing dots on a whiteboard with lines between them (Gleicher 2018). But even this lackluster description remains one of the most detailed a platform has given, for any of the categories meant to capture information operations.[20] Facebook also, in October 2020, released a report on "Inauthentic Behavior," the overarching category of which CIB is a subset (Gleicher 2020). It presented this inaugural report as the first in a series and noted, "our goal with this new reporting series is to share trends and tactics we see in IB. … By publicizing our findings, we aim to advance the public's understanding of this evolving space, including the gray areas where harm and deception aren't as clear cut" (ibid.).

This appeared to be a welcome step forward in taking the practice routinely applied to CIB we are describing in this article and expanding the scope of that pocket of transparency.

18. See, influentially, Wardle and Derakhshan (2017); see also François (2019).
19. See, e.g., Gadde and Roth (2018).
20. See "The Lawfare Podcast: Alex Stamos on the Hard Tradeoffs of the Internet" (2020) describing how Facebook was going to call the category "Coordinated Inauthentic Activity," but decided "CIB" made a better acronym than "CIA."

Facebook explained, "in future reports, we will share more examples of these gray area behaviors and how we tackle them" (Gleicher 2020). But there were no future reports. A year later, the inaugural *Inauthentic Behavior* report stands as the first and only one in this "series."

Twitter's approach to what it decides to disclose in its "Information Operation archive" differs from Facebook's "CIB disclosures" in important ways. Twitter's public archive of "information operations" only includes accounts that participated in "[p]latform manipulation that [can reliably be attributed] to a government or state linked actor" ("Information Operations" 2018). Twitter's definition of "platform manipulation" is detailed ("Platform manipulation and spam policy" 2020), but delimiting its archive to accounts that can reliably be attributed to state actors is a significant limiting factor for scope. It is also a departure from the behavior-centric definition used by Facebook, by approaching the subset of platform manipulation it considers "Information Operations" for inclusion in the archive in an actor-centric way. What differentiates "routine" platform manipulation from information operations is the platform's ability to attribute the activity to a government or a state-linked actor. Attribution therefore becomes the key criterion for inclusion in the Information Operations archive; a campaign designed to manipulate public opinion but which cannot be reliably attributed "to a government or state linked actor" would presumably be enforced against (i.e., participating accounts would be suspended), but not included in the Information Operations archive. This criteria puts a lot of pressure on attribution, which in itself is a nuanced, evolving, and difficult process.[21] Twitter has publicly acknowledged the difficulties of attribution: in February 2019, it updated its January 2018 IRA disclosure, noting that 228 accounts had been "misidentified as connected to Russia." The update to the update notes that "additional information" allowed Twitter to "more confidently associate [the accounts] with Venezuela."[22] It is also important to note that Twitter's archive is the most thorough within its scope: once a campaign is selected for inclusion in the archive, Twitter provides the most comprehensive data to the public out of all platforms. Twitter's focus on state-linked campaigns, while *limiting* in scope, is *enabling* in breadth: accounts run by state-based actors, rather than non-state actors or other individuals, involve different privacy tradeoffs and may enable a more comprehensive sharing of data and metadata associated with the account(s). But at a time when an increasingly wide variety of actors engage in on-platform information operations, this can seem like an unduly narrow view of information operations.

Google's policy invokes different terminology yet again, announcing that its quarterly bulletin would cover "coordinated influence operation campaigns" (Huntley 2020a). It has never provided a precise definition of what these "COICs" include, or how Google determines what activity meets this threshold. The May 2020 announcement that accompanied the inaugural bulletin noted that the purpose of the quarterly document would be to "share information about actions we take against accounts that we attribute to coordinated influence campaigns (foreign and domestic)" (Huntley 2020b). Looking at the patterns emerging from the operations disclosed by Google to date, one can conclude that Google's disclosure criteria is broader than Twitter's *actor-centric* criteria, as it does include campaigns attributed to non-state actors entities, but seemingly narrower than Facebook's CIB *behavior-centric* disclosure criteria (given that CIB networks disclosed by Facebook and with corresponding assets on Google platforms seemingly do not meet the threshold for inclusion).

Reddit uses the umbrella of "content manipulation" to define the types of campaigns

21. For discussion of the difficulties and nuances of attribution in cybersecurity, see Rid and Buchanan (2015) and Egloff and Smeets (2021). How these cybersecurity attribution principles hold (and do not hold) when applied to information operations remains basically unexplored in the literature.

22. An endnote to this policy post points to a Twitter thread by Twitter's head of Site Integrity, Yoel Roth, crediting an independent researcher for pointing the initial inconsistency (Roth 2019).

that violate its rules, which it vaguely describes as "a term we use to combine things like spam, community interference, etc." (u/worstnerd 2019). Only a handful of campaigns have ever been disclosed by Reddit; it is unclear if the platform has taken a commitment to disclose *all* information operations it has ever identified on its platforms despite only having found a handful or if more campaigns were found that were not disclosed.

As a Carnegie Endowment analysis of these and nine other platforms' broader content moderation policies concludes, the terminology and substance platforms use in their community standards to describe information operations and related manipulation "varies significantly across platforms" (Bateman et al. 2021). As a result, the policies on paper do not give a good picture of how much overlap and divergence there is in what platforms take down.

The disclosures platforms make of how they have enforced their policies provide more insight into how much overlap there is (or is not) in what different platforms' policies cover. That is, one can partially "reverse engineer" the contours of the policies to see what falls under their umbrellas. There is some remarkable and counterintuitive divergence, some of which is more apparent than others. For instance, Facebook routinely discloses domestic campaigns as part of their CIB reports and has notably disclosed a handful of US-based CIB operations.[23] Platforms often reference information disclosed by Facebook as a lead for finding related activity on their services, and Facebook systematically notes in their monthly CIB reports that it shares information about its findings with industry partners. But *none* of the US-based campaigns disclosed by Facebook have ever appeared in corresponding disclosures from Twitter or Google, even when reporting has confirmed that these campaigns spread across multiple platforms.[24] Good examples are a US-based domestic CIB campaign linked to Turning Point USA and a US-based campaign linked to Roger Stone and associates, both taken down by Facebook and spreading across platforms, but not included in Google's CIOC or Twitter's IO disclosures (in the case of Twitter, because the campaign did not involve a state actor) (Graphika Team 2020; Stanley-Becker 2020).

Similarly, independent researchers have repeatedly exposed campaigns from a China-based large scale, cross-platform political influence operation dubbed "Spamouflage" for its use of classically spammy techniques, such as mass account creation and low-quality, high-volume content production.[25]. These exposures have in turn triggered multiple waves of takedowns across platforms.[26] The disclosures related to these takedowns appear in Google's quarterly bulletin,[27] suggesting that Spamouflage's campaigns constitute CIOC. But they do not appear in Facebook's CIB reports, suggesting—without explanation—that these campaigns don't meet the policy threshold for Facebook. And despite no industry consensus (or public acknowledgment) that these operations can reliably be traced back to a government or a state-linked actor, data from these takedowns is included in Twitter's Information Operations archive.

There are many possible reasons for these discrepancies in disclosures across the industry. Sometimes it may simply be that the campaign in question targets only one platform (although it is worth noting that researchers find these operations are increasingly cross-

23. In this context, "domestic" refers to operations originating in one country and targeting the population of that same country.
24. This is true as at the time of publication.
25. See, e.g., Nimmo et al. (2021).
26. See, e.g., Timberg and Harris (2021).
27. See for instance TAG Bulletin from Q2 2020, published August 5, 2020: "We terminated 186 YouTube channels as part of our ongoing investigation into coordinated influence operations linked to China. These channels mostly uploaded spammy, non-political content, but a small subset posted political content primarily in Chinese similar to the findings in a recent Graphika report, including content related to the U.S. response to COVID-19" ("TAG Bulletin" 2020).

platform).[28] But a campaign that exists across two platforms could also be classified differently by each platform, and so meet the disclosure criteria for one platform and not another. It could also reflect a timing discrepancy, given the priorities of the internal enforcement team at that time, with resourcing or other constraints leading platforms to act at different times or not at all based on their own internal assessments of urgency. It may be the case that a particular campaign manifests differently on different platforms, with activity that would amount to information operations or similar manipulative behavior on one platform but not on another, simply because of the way the particular actors exploit different platforms' affordances. A good example is one described by Reddit in October 2019:

> Let's look at what other major platforms have reported on coordinated behavior targeting Hong Kong. Their investigations revealed attempts consisting primarily of very low quality propaganda. This is important when looking for similar efforts on Reddit. In healthier communities like r/hongkong, we simply don't see a proliferation of this low-quality content (from users or adversaries). The story does change when looking at r/sino or r/Hong_Kong (note the mod overlap). In these subreddits, we see far more low quality and one-sided content. However, this is not against our rules, and indeed it is not even particularly unusual to see one-sided viewpoints in some geographically specific subreddits. ... What IS against the rules is coordinated action (state sponsored or otherwise). We have looked closely at these subreddits and we have found no indicators of widespread coordination. In other words, we do see this low quality content in these subreddits, but it seems to be happening in a genuine way. (u/KeyserSosa 2019)

The simplest explanation is, of course, that platforms' policies are ambiguous and simply cover different kinds of activity. Decisions about what constitutes illegitimate platform manipulation are not objective, technical determinations. They require platforms to make subjective judgments as to whether certain operations meet their thresholds and (as a factor in this determination) whether they want to publicly announce what they found.

This dynamic is the key problem we seek to highlight in this paper: the transparency regimes that were the result of unprecedented public pressure following 2016 have upstream consequences for platforms' policy definitions and enforcement activities. Because there is ambiguity and flexibility in defining "information operations" or "platform manipulation," a platform's calculus about the costs and benefits of a particular disclosure could influence how it defines that category. And because information asymmetries between platforms and external observers are especially acute in the context of platform manipulation (François 2019), it is impossible for outsiders to know how platforms make these determinations.

There is one further complication that compounds this opacity. A CIB/IO/CIOC classification isn't necessarily an exclusive one. A campaign can—and often will—constitute CIB/IO/CIOC and another policy violation, such as spam, hate speech, medical misinformation, or harassment, for example. But different kinds of policy violations are subject to different (and less detailed) disclosure regimes and tend to be detected by different teams using different systems. Platforms do not publicly disclose when there is overlap, how they determine which rule they will enforce under, and how the different transparency regimes influence this decision. This "closest cousins" problem—the fact that information operations often border or completely overlap with other content moderation categories—is a key reason why these disclosure reports should not be considered

28. See, for instance, "The State of Influence Operations 2017–2020" (2020, 26).

as complete records of all these operations on the platforms. They are simply records of the campaigns these platforms chose to investigate and disclose under a specific policy at a given time. As we will discuss in Section 6, the "closest cousins" problem is a growing one. Platform policies are evolving to address a greater array of online abuses and generating new overlapping and opaque categories, with inconsistent disclosure standards.

To summarize, platform disclosures of information operations are underinclusive for a number of reasons. First, platforms may simply not enforce against certain types of activity because of internal constraints, operational priorities, or commercial considerations, or because they decide that the activity does not meet their circumscribed definition of information operations for some other reason. Reporting and independent analysis of platform disclosures, for example, suggests that there are geographical- and public relations-driven biases in what campaigns platforms prioritize investigating and taking action against.[29] Second, certain campaigns will often be taken down under other content moderation categories, like hate speech or spam, either as a conscious categorization decision by platforms to avoid the more extensive disclosure obligations that attend taking action under a CIB/IO/CIOC policy or simply as a result of automated moderation tools or other content moderation processes removing such activity before it attracts the attention of the team in charge of detecting and enforcing against CIB/IO/CIOC. Third, as evidenced by Twitter's focus on state-linked campaigns or Google's lack of US-based domestic campaigns in their archives, the categories themselves are more limited than the general, colloquial understanding.

5 The Legacy of 2016: A Frozen and Distorted Regime

This is where we stand, then. The discovery of 2016 Russian interference led to a beneficial and meaningful set of transparency regimes that provide some insight into how platforms police this kind of activity but are still flawed, increasingly out of date, and not always maintained by platforms. These regimes were a response to a particular political moment and a particular instance of platform manipulation by Russian actors, and that moment still casts a long shadow over the field of influence operations in general. Until remarkably recently, this regime created a disproportionate focus on foreign influence campaigns at the cost of ignoring the very real and often more problematic effects of domestic influence campaigns.[30] It has also stifled the analytical work of probing the hard questions of what constitutes improper online activity more generally.[31] The colloquial usage of "CIB" illustrates that it has become a generic stand-in for "problematic online behavior," despite it being a limited category, confined to one platform, which obscures the definitional work that still needs to be done to describe the range of manipulative behavior online and the variance in policies across companies.

But the most important distortion is the disclosure incentives the regimes create. For the team responsible within a platform, the stakes are high as to whether platform manipulation is classified as CIB/IO/CIOC: on one side of the line is the most comprehensive kind of platform disclosure that currently exists with specific descriptions and sometimes data as to what was enforced against. On the other side of the line is limited, if any, separate disclosure of platform detection or action. But the line is fuzzy and, at times, illusory. The fuzziness and difficulty of defining what constitutes problematic online behavior has been discussed by many scholars, including ourselves (François 2019; douek 2020). Our point here is related but different: arbitrary line-drawing, overlapping categories, or disparities

29. See, e.g., Silverman et al. (2021) and Wong (2021a, 2021b, 2021c, 2021d, 2021e).
30. See, e.g., DiResta (2020), Wardle (2020), and Benkler (2020).
31. See, e.g., Lim (2020), DiResta (2020), Benkler (2020), Wardle (2020), and douek (2020).

across platforms would not be so problematic if the transparency outcomes were not so radically different. For example, assets (accounts, posts, ads, etc.) in a campaign may be publicly disclosed, provided to independent researchers or to the public for further analysis if classified as CIB/IO/CIOC (Election Integrity Partnership 2021, 106ff). But if a group of assets are taken down as "spam," then the enforcement numbers just get added to the large aggregated figures platforms release in their separate terms of service enforcement transparency reports.

The fuzzy and shifting borders that CIB/IO/CIOC-style behaviors share with many other categories of platform manipulation or abuse create an opening for many to exploit these ambiguous and opaque boundaries, including the platforms themselves. As discussed, in some cases, this will incentivize under-classifying activity as CIB/IO/CIOC, to avoid the extra attention and scrutiny that accompanies transparency. In other cases, it may incentivize over-classification, to demonstrate how very seriously platforms are taking these problems and how hard they are working to root out this kind of abuse. Platforms largely received a positive report card for such disclosures around the 2020 US election, for example, by showing that they had uncovered and removed more information operations and more quickly, before they were able to gain significant traction online.[32] But without knowing the denominator of how many campaigns existed or whether platforms were simply choosing to classify more activity as information operations for the purposes of disclosure, it is difficult to reliably use this data to deduce overall trends within the field.

Researchers may also have an incentive to argue that campaigns that are identified as harmful are classified as CIB/IO/CIOC, as this currently is the only classification regime that would lead to acknowledgment and additional publicity regarding their work.

Finally, governments also participate in sharing "leads" with platforms regarding these types of operations. All major platforms have acknowledged receiving leads from the US government (including in the context of their cooperation to secure the 2020 US presidential election), leading to investigations and enforcement action against campaigns classified as CIB/IO/CIOC. Another example from France demonstrates some reasons why these relationships may raise concerns. In May 2021, Facebook disclosed a CIB network run by "individuals associated with the French military" targeting audiences in Mali and the Central African Republic. The French network seemed to focus on Russian efforts targeting Francophone audiences in these countries and criticizing the French government. Shortly after Facebook's disclosure and the publication of a companion report by Graphika and the Stanford Internet Observatory (Graphika & SIO 2020), Politico revealed that the French government had itself sent emails to Facebook "outlining evidence of Russia's disinformation campaign on social media across Africa" (Scott and Braun 2020). That is, France was running its own campaigns in Africa while simultaneously giving tips to Facebook about Russia engaging in similar behavior, presumably with the goal of having the Russian campaigns taken down. These emails had not been made public prior to Politico's reporting. This raises concerns that governments will use the vagueness of the CIB/IO/CIOC category and the unusually close relationships they have with platforms in this context for their own political and strategic ends. The relationships between platforms and governments are especially opaque in this context because platforms act on informal tips from governments, not legal orders, and ostensibly rely on platforms' own determination of the content as violating their standards. Thus, platforms' enforcement actions taken as a result of government tips would not be included in the separate reports they release about government takedown requests,

32. See, e.g., Newton (2021): "Reading through the report, there's a lot to be impressed by. Foreign interference, which all but defined the 2016 US presidential election, played almost no perceptible role in 2020. After making huge investments in safety and security, platforms really did get better at identifying fake accounts and state-backed influence campaigns, and generally removed them before they could do much about them."

which typically only include formal legal orders. Governments have their own incentives in leveraging these disclosures for deterrence, playing a game of "name and shame" by proxy.

Disinformation actors themselves will also attempt to take advantage of CIB/IO/CIOC disclosure regimes. Indeed, some of these actors may seek publicity for their work, and the goal of specific campaigns can be to yield public coverage[33]—a strategy referred to as "perception hacking." The Internet Research Agency's campaign to target the 2018 US midterms is an example: the operation appeared to have been designed to be exposed, with messages boasting about the troll farm's ability to continue manipulating American voters despite the defensive efforts put in place since 2017. A manifesto published online as part of the operation read, "Soon after November 6, you will realize that your vote means nothing. We decide who you vote for and what candidates will win or lose. Whether you vote or not, there is no difference as we control the voting and counting systems. Remember, your vote has zero value. We are choosing for you" (Foer 2020). In short, the ambiguous pocket of transparency created by the CIB/IO/CIOC can also act as a magnet for certain actors seeking to have their manipulative efforts publicly disclosed and discussed as part of their strategy to self-aggrandize.

All these dynamics operate against a background of perverse transparency incentives for platforms generally. Often, the more platforms disclose, the more they can be scrutinized and criticized. When all disclosures are voluntary, platforms that *do* decide to disclose often receive more attention and blame than those that don't, regardless of whether its services host more manipulative behavior than others. Take Facebook and Twitter's ban of Turning Point USA and related parties, for example—this prompted criticisms about why those platforms had not acted sooner or more comprehensively (Stanley-Becker 2020). Meanwhile, Google did not disclose if it had even investigated related activity on its services but did not receive criticism for not doing so. More broadly, many platforms—including increasingly important ones like TikTok—have vague standards on information operations and make no disclosures whatsoever regarding campaigns and content that has been removed under these standards. Others, like Parler or Gab, refuse to remove information operations even once they have been found and reported to them by independent researchers (Timberg 2020). Despite all this, most of the scrutiny tends to fall on Facebook and Twitter, the platforms that tend to disclose the most.

The real question here should be "what CIB/IO/CIOC campaigns are not being disclosed, and why?" more than "which platform discloses the most?" This state of affairs will inevitably lead to further tensions, unless platforms either commit to be comprehensive in their disclosures of all campaigns that meet their threshold, or to be more forthcoming on the reasons why some are not included in these special CIB/IO/CIOC reports.

6 Closest Cousins

The issues raised by the ambiguous scope of CIB/IO/CIOC policies and the incentives created by the unique transparency regimes for this kind of content moderation are becoming more acute. There are an increasing number of what this article calls "closest cousins" categories of content moderation—other types of online behavioral content moderation that share a border with CIB/IO/CIOC-style operations but without the same transparency regimes.

Precisely because these similar behavioral content moderation categories do not have the same disclosure practices as CIB/IO/CIOC enforcements, assessing how different

33. An infamous instance of this "perception hack" strategy is the Russian Internet Research Agency's campaign around the US 2018 midterms; see Glaser (2018).

platforms' definitions and policies overlap is difficult. Therefore, our goal in this section is not to provide a comprehensive review of platform policies, or to make an argument for where substantive lines should be drawn. Instead, the objective of this section is to provide examples that illustrate how the growing number of content categories that brush up against or overlap with CIB/IO/CIOC has underappreciated consequences given the very different transparencies that platforms provide into their moderation of each of these categories. Some of these categories predate the creation of CIB/IO/CIOC content moderation, others postdate it. What they have in common is that they do not focus on *content* as the determinant of whether a rule has been violated. But unlike CIB/IO/CIOC, there are no specialized transparency regimes that shed light on how platforms are drawing the precise lines around these categories.

Spam is one of the original forms of content moderation—no platform can survive without a strategy for dealing with spam—and such moderation often relies on signals other than the content of individual posts. Spam is one of the largest categories of content moderation that platforms engage in, in terms of volume, but also a category that attracts little public attention (Jeong 2018). However, "the definition of spam is nebulous" (ibid.) and implicates often underappreciated value judgments about what constitutes acceptable online behavior.[34] The "Spamouflage" campaign discussed above demonstrates the thinness of the line between how platforms define CIB/IO/CIOC and how they define spam. Another example is the decision by YouTube, in the heat of the 2020 election, to remove some highly viewed YouTube videos broadcasting fake election results on Election Day on the grounds that they were spam (the platform had not released a policy against false claims of election victory, as other platforms had [Tenbarge 2020]). As is typical of spam takedowns, there was no explanation for the decision. These examples highlight that some of the most fraught and politically significant content moderation decisions can be categorized as spam. When or how often this occurs is unknown. This creates a dynamic that goes underappreciated: complete opacity on the boundaries of the category "spam" means that platforms can classify the same content as an information operation *or* spam depending on their own transparency incentives. If they classify the takedown as the former, a comprehensive disclosure is expected; if the latter, there is no meaningful transparency at all.

This equally applies to the vast array of "inauthentic" or "coordinated" online activity that does not meet a platforms' particular definition of CIB/IO/CIOC. As Tarleton Gillespie has observed, the distinction between "coordinated efforts" to game a system and the "genuine" output of users is a false one: "[m]ost contributions to the web are somewhere in the middle, where people in some way coordinate their efforts in order to help make their content visible to a search engine, out of a 'genuine' desire for it to be seen" (Gillespie 2017). The socially constructed nature of online "authenticity" and its fluidity and contextual nature has been a long-standing topic in social media scholarship.[35]

In October 2020, Facebook released its first report on its version of this kind of "inauthentic behavior" (IB), the umbrella policy under which CIB sits (Gleicher 2020). The report shows how hard defining this category can be, noting "both legitimate, highly active users and deceptive actors regularly develop new techniques that test the boundaries of our policies" ("Inauthentic Behavior Report" 2020, 3). Who is a "legitimate, highly active user" and who is a "deceptive actor" raises fraught questions without objective or determinate answers. The line between good faith grassroots political campaigning or aggressive marketing and illegitimate manipulation can be a fine one. Twitter also releases a report on its "platform manipulation" takedowns that do not constitute "information operations," ("Platform Manipulation," n.d.) and the difference in the level

34. On the broadness and ambiguity of the definition of "spam," see Brunton (2013).
35. See, foundationally, Marwick and Boyd (2011).

of transparency provided compared to its public archive of "information operations" is stark. These reports only include aggregate figures of spam taken down per six-month period.

Major platforms are also unveiling an increasing array of policies intended to address when online real users coordinate to cause online or offline "harm" beyond informational harms or operations. Twitter has unveiled a "coordinated harmful activity" policy that includes a detailed matrix for decision making along three different vectors and extended narrative descriptions of the different criteria for classification along each vector ("Coordinated Harmful Activity," n.d.). But despite having been released in January 2021 and the platform invoking it a number of times since then to justify high profile takedowns—most notably, a harassment campaign directed at the so-called Queen of Twitter (douek 2020b), Chrissy Teigen—Twitter has only ever referred to the policy at a high level and has never explained where in the detailed matrix any particular enforcement action falls, leaving its decision-making processes obscure.

Facebook has also unveiled a new "coordinated social harm" policy, designed to target the coordination between *authentic* users that the CIB policy does not address (Gleicher 2021). But the blog post announcing the new policy is vague and the policy itself has not yet been added to Facebook's Community Standards. Therefore, as well as raising all the same definitional issues around what constitutes "coordination" that the CIB policy raises, it introduces new ones about how Facebook defines "social" and "harm" that remain unaddressed. Facebook has only enforced the policy once so far—against a German movement called Querdenken. There is therefore no pattern of enforcement that would fill in the sparse details of the new policy. The same vagueness issues apply to Facebook's new "Coordinated Mass Harassment" policy under which it will remove certain coordinated efforts to harass others "even if the content on its own wouldn't violate our policies," but how Facebook will decide when those efforts cross the line into this new policy domain is unclear (Davis 2021).

As these categories expand, platforms continue to diverge in their approach to dealing with the same underlying activity. QAnon provides an especially relevant example. In a moment reminiscent of how platforms responded to public and regulatory pressure about their handling of election interference in 2016, platforms took action in 2020 against QAnon activity on their services after months of public pressure to do so. But they justified that action in different ways. Twitter focused on the *behavior* of the accounts involved, taking down tens of thousands of accounts under its coordinated harmful activity policy (Herridge et al. 2021). Facebook focused on the *actors* behind that activity, expanding its Dangerous Individuals and Organizations policy to cover "militarized social movements" (Facebook 2020). YouTube focused on the *content*, expanding its hate and harassment policies to prohibit "content that targets an individual or group with conspiracy theories that have been used to justify real-world violence" (Team 2020). Many other platforms took action in response to public pressure, falling like dominoes, with little to no explanation of why or whether this would be a consistent new approach (Gonzalez 2021). The pressure was so intense that even fitness platform and exercise bike company Peloton banned QAnon content (Oremus 2020).

Platforms' policies against these coordinated efforts that they deem "harmful" and requiring removal are therefore the perfect illustration of the issues these expanding "closest cousins" categories raise. Platforms are increasingly turning to content moderation based on criteria other than content to address a number of identified harms created through their services, but the boundaries of these categories are vague and the transparency they provide is inconsistent and varies based on a particular platforms' approach. What *is* consistent across the board, though, is that the transparency that accompanies the most high-profile category of behavioral content moderation (CIB/IO/CIOC) is lacking

(Newton 2020).

6.1 The Transparency Cliff

As this brief sketch should make clear, there are many closely related categories of content moderation that are adjacent or overlapping with CIB/IO/CIOC content moderation. But by contrast with CIB/IO/CIOC enforcement actions, there is little to no transparency or explanation when platforms act against these kinds of user behavior. Through a sheer accident of content moderation history, other categories of content moderation do not fit within the small transparency spotlight that the aftermath of 2016 created. Efforts to expand that spotlight, such as Facebook's inaugural *Inauthentic Behavior* report, have seemingly been abandoned.

To be clear, this is not to say that no transparency exists in other areas. The idea of "platform transparency" is broad and can cover many things (Gorwa and Garton Ash 2020, 286, 294–95; Ausloos, Leerssen, Thije, et al. 2020). In some areas, platforms have offered more long-standing and sometimes extensive transparency, such as reports to the Lumen database of pieces of content taken down based on legal orders (Lumen 2019), or the growing practice of platforms voluntarily providing archives of advertisements run on their sites (Leerssen et al. 2018). Years of advocacy outside and inside platforms, investigative reporting, and efforts such as the *Santa Clara Principles on Transparency and Accountability in Content Moderation* have also brought more transparency around platforms' rules and voluntary reports of content takedowns under their terms of service generally.[36] Google pioneered this practice in 2010, and most of the major platforms now regularly release high-level overviews of their enforcement efforts. These reports have slowly expanded and included more detail over time. Twitter, for example, now provides around 10 transparency reports, whose coverage includes "Rules Enforcement," "COVID 19 Misinformation," and the closest cousin "Platform Manipulation," as noted above. Google and Facebook provide quarterly reports covering their enforcement of their community standards generally across their services, including YouTube for Google and Instagram for Facebook. This genre of report, with varying levels of detail and frequency, is becoming standard across the social media platform industry. Separate reports for enforcement actions in response to legal obligations such as intellectual property rights, government reports of content violating local law, and law enforcement requests for user data are also now common.[37]

But while these reports provide aggregate figures and can show general trends, they do not contain the level of detail or specific examples included in CIB/IO/CIOC reports. In the realm of voluntary takedowns under platforms' terms of service, the CIB/IO/CIOC transparency regime remains unique in the level of detail it offers. The disparity is vast, and what makes any given takedown fall into one of the most transparent categories of content moderation or one of the most opaque is often entirely unclear.

36. See, e.g., Wong and Solon (2018), Grassegger and Angwin (2017), and Santa Clara Principles (n.d.).

37. For a thoughtful discussion of how these broader platform transparency reports evolved over the past five years, see Brouillette (2020). In this study, authors note that Twitter's Rules Enforcement report, along with others that the platform publishes, stands out in the industry "for being far more transparent about its content moderation practices than any other platform we ran."

7 Conclusion

Calls for transparency in the ongoing content moderation debates are as common as they are often vague (Suzor et al. 2019; Keller 2021). In among this milieu, the CIB/IO/CIOC regime exhibits a number of distinctive features: separate internal teams dedicated to monitoring and enforcing against the relevant on-platform activity; the release of regular public reports and announcements of platforms' enforcement measures; provision of underlying data to external stakeholders and, in Twitter's case, the public; and cross-industry collaboration and tip-offs to other (including smaller) platforms. Even as there remain valid criticisms about the form, extent, and sometimes selective nature of CIB/IO/CIOC disclosures, they have also demonstrated the benefits of transparency. They have enabled accountability journalism (Legum 2019), public reporting on the ongoing nature of information campaigns,[38] and academic research into the nature and extent of influence operations,[39] as well as discussion around the need for definitional clarity in how platforms define these activities (François 2019; douek 2020c; McGregor 2020). The spotlight these disclosures have helped create is especially valuable when it comes to platforms' global markets, which platforms have consistently neglected. And when most of these campaigns operate across platforms, disclosures also help create incentives for platforms to find and remove such activity and collaborate through information sharing, lest they appear negligent compared to their peers.

But in providing a limited window of transparency—and an arbitrary one at that—the CIB/IO/CIOC spotlight has also created an outsized focus on this kind of information operation at the expense of examining the vast corpus of "closest cousin" online manipulative behaviors. The CIB/IO/CIOC disclosure regime has spawned a regular and predictable media cycle: platforms detect a campaign, they disclose data to trusted partners who write reports describing what platforms found, and the media covers these reports with exciting and attention-grabbing headlines about information operations. The fact that such campaigns no longer gain much traction is often lost. The public perception of content moderation and how the online public sphere is manipulated is skewed as a result toward a very limited slice of online influence operation, which increasingly plays an ever-smaller role in how actors manipulate the online information ecosystem and the responses platforms are engaging in to counter these different threats. The outsized emphasis on CIB/IO/CIOC is exacerbated by the fact that the limited and arbitrary nature of these transparency regimes remains underappreciated. This article has sought to highlight these features by recounting the contingent nature of how these regimes came into being. Far from a planned out and deliberate project from the start, CIB/IO/CIOC reporting was the product of a particular political moment. A failure to understand the distorting effects this has had means that these transparency regimes threaten to obscure more than they illuminate.

Finally, the early promise and momentum behind the creation of these pockets of transparency are being lost as public and regulatory focus and pressure turns to other areas of content moderation. There are signs that platforms' commitment to and ongoing investment in public transparency in this domain are decreasing, just as they increase the kinds of content moderation that borders that very domain. This risks leaving these unique pockets of transparency as an increasingly obsolete accident of history—an unfinished sentence that promised a new era of accountability that will not materialize.

38. See, e.g., Isaac (2019), Conger (2019), and Romm (2020).
39. Most notably, of course, SSCI (2018), DiResta et al. (2019), and Howard et al. (2019). See also Colliver et al. (2020).

References

Alba, Davey. 2020. "Facebook Removes Roger Stone for Ties to Fake Accounts," https://www.nytimes.com/2020/07/08/technology/roger-stone-facebook.html.

Ausloos, Jef, Paddy Leerssen, Pim ten Thije, et al. 2020. "Operationalizing Research Access in Platform Governance What to learn from other industries?," https://www.ivir.nl/publicaties/download/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf.

Bateman, Jon, et al. 2021. "How Social Media Platforms' Community Standards Address Influence Operations." *Carnegie Endowment for International Peace,* accessed April 15, 2021. https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201.

Benkler, Yochai. 2020. "The Danger of Overstating the Impact of Information Operations." *Lawfare,* accessed November 22, 2020. https://www.lawfareblog.com/danger-overstating-impact-information-operations.

Berger, J. M. 2015. "The Evolution of Terrorist Propaganda: The Paris Attack and Social Media," https://www.brookings.edu/testimonies/the-evolution-of-terrorist-propaganda-the-paris-attack-and-social-media/.

Bickert, Monika, and Brian Fishman. 2017. "Hard Questions: How We Counter Terrorism." Facebook Newsroom (Jun. 15, 2017), https://about.fb.com/news/2017/06/how-we-counter-terrorism/.

Brouillette, Amy. 2020. "Key Findings," https://rankingdigitalrights.org/index2020/key-findings.

Brunton, Finn. 2013. *Spam—A Shadow History of the Internet.* The MIT Press.

Clark, Mike. 2021. "Research Cannot Be the Justification for Compromising People's Privacy." *Facebook Newsroom,* https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/.

Colliver, Chloe, et al. 2020. "Hoodwinked: Coordinated Inauthentic Behaviour on Facebook."

Conger, Kate. 2019. "Facebook and Twitter Say China Is Spreading Disinformation in Hong Kong." *New York Times,* accessed April 17, 2021. https://www.nytimes.com/2019/08/19/technology/hong-kong-protests-china-disinformation-facebook-twitter.html.

Davis, Antigone. 2021. "Advancing Our Policies on Online Bullying and Harassment." *Facebook Newsroom,* https://about.fb.com/news/2021/10/advancing-online-bullying-harassment-policies/.

Department of Justice. 2018. "Grand Jury Indicts Thirteen Russian Individuals and Three Russian Companies for Scheme to Interfere in the United States Political System," accessed April 14, 2021. https://www.justice.gov/opa/pr/grand-jury-indicts-thirteen-russian-individuals-and-three-russian-companies-scheme-interfere.

DiResta, Renée. 2020. "The Conspiracies Are Coming From Inside the House." *The Atlantic,* https://www.theatlantic.com/ideas/archive/2020/03/internet-conspiracies-are-coming-inside-country/607645/.

DiResta, Renée, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. "The tactics & tropes of the Internet Research Agency."

douek, evelyn. 2020a. "The Rise of Content Cartels." *Knight First Amendment Institute at Columbia University,* accessed February 14, 2020. https://knightcolumbia.org/content/the-rise-of-content-cartels.

———. 2020b. "Twitter Brings Down the Banhammer on QAnon." *Lawfare,* https://www.lawfareblog.com/twitter-brings-down-banhammer-qanon.

———. 2020c. "What Do Platforms Think "Coordinated Inauthentic Behavior" Actually Means?" *Slate,* accessed August 31, 2020. https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html.

Dwoskin, Elizabeth. 2017. "How Russian Content Ended up on Pinterest," accessed April 14, 2021. https://www.washingtonpost.com/news/the-switch/wp/2017/10/11/how-russian-content-ended-up-on-pinterest/.

Economist, The. 2018. "The Techlash against Amazon, Facebook and Google—and What They Can Do." *The Economist,* accessed January 12, 2019. https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do.

Egloff, Florian J., and Max Smeets. 2021. "Publicly Attributing Cyber Attacks: A Framework." *Journal of Strategic Studies.*

Election Integrity Partnership. 2021. *The Long Fuse: Misinformation and the 2020 Election.* V1.3.0. Stanford Digital Repository: Election Integrity Partnership. https://purl.stanford.edu/tr171zs0069.

evelyn douek. 2020. "The Free Speech Blind Spot: Foreign Election Interference on Social Media." In *Defending Democracies,* 265–292. Oxford University Press.

Facebook. 2020. "An Update to How We Address Movements and Organizations Tied to Violence." *Facebook Newsroom,* https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

"Community Standards: 22. Inauthentic Behaviour." 2020. *Facebook Community Standards,* accessed January 23, 2020. https://www.facebook.com/communitystandards/inauthentic_behavior.

"Removing Bad Actors on Facebook." 2018. *Facebook Newsroom,* https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/.

Foer, Franklin. 2020. "Putin Is Well on His Way to Stealing the Next Election," https://www.theatlantic.com/magazine/archive/2020/06/putin-american-democracy/610570/.

Fogel, Mikhaila. 2018. "Documents: Senate Intelligence Committee Publishes Two Reports on Internet Research Agency." *Lawfare,* accessed April 14, 2021. https://www.lawfareblog.com/documents-senate-intelligence-committee-publishes-two-reports-internet-research-agency.

Foroohar, Rana. 2018. "Year in a Word: Techlash." *Financial Times,* accessed January 12, 2019. https://www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e.

François, Camille. 2019. "Actors, Behaviors, Content: A Disinformation ABC, Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression," https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf.

François, Camille, and Herbert Lin. 2020. "The strategic surprise of Russian information operations on social media in 2016: Mapping a blind spot." *Herodote,* no. 2, 33–57.

Gadde, Vijaya, and Yoel Roth. 2018. "Enabling Further Research of Information Operations on Twitter." *Twitter Blog,* accessed April 14, 2021. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html.

Gillespie, Tarleton. 2017. "Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem." *Information, communication & society* 20 (1): 63–80.

Glaser, April. 2018. "Was the Latest Online Propaganda Campaign Busted by Facebook the Work of Russian Trolls—or Trolls Impersonating Russian Trolls?" *Slate Magazine,* accessed May 15, 2021. https://slate.com/technology/2018/11/facebook-ira-russia-trolls-midterms.html.

Gleicher, Nathaniel. 2018. "Coordinated Inauthentic Behavior Explained." *Facebook Newsroom,* https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

———. 2020. "Inauthentic Behavior Report: Update on Our Work Against Deceptive Behavior." *Facebook Newsroom,* https://about.fb.com/news/2020/10/inauthentic-behavior-report.

———. 2021. "Removing New Types of Harmful Networks." *Facebook Newsroom,* https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/.

Gonzalez, Oscar. 2021. "QAnon FAQ: Who is Q and What You Should Know about this pro-Trump Conspiracy Theory." *CNET,* https://www.cnet.com/news/qanon-faq-who-is-q-and-what-you-should-know-about-this-pro-trump-conspiracy-theory/.

"TAG Bulletin: Q2 2020." 2020. *Google,* https://blog.google/threat-analysis-group/tag-bulletin-q2-2020/.

Gorwa, Robert, and Timothy Garton Ash. 2020. "Democratic Transparency in the Platform Society." *Social Media and Democracy: The State of the Field, Prospects for Reform.*

Graphika & SIO. 2020. "More-Troll Kombat," https://public-assets.graphika.com/reports/graphika_stanford_report_more_troll_kombat.pdf.

Graphika Team. 2020. "Facebook's Roger Stone Takedown: Facebook Removes Inauthentic Network Attributed to Political Operative."

Grassegger, Hannes, and Julia Angwin. 2017. "Facebook's Secret Censorship Rules Protect White Men…" *ProPublica,* accessed December 14, 2018. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Grimmelmann, James. 2015. "The Virtues of Moderation." *Yale Journal of Law & Technology* 17 (42): 47.

Herridge, Catherine, et al. 2021. "After Years of Trying to Curb QAnon Messaging, Twitter Has Now Suspended More than 150,000 Accounts," accessed April 2, 2021. https://www.cbsnews.com/news/qanon-twitter-suspends-150000-accounts-capitol-riot/.

Howard, Philip N, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2019. "The IRA, social media and political polarization in the United States, 2012–2018."

Huntley, Shane. 2020a. "TAG Bulletin: Q4 2020." *Google,* https://blog.google/threat-analysis-group/tag-bulletin-q4-2020/.

———. 2020b. "Updates about Government-Backed Hacking and Disinformation." *Google,* accessed April 14, 2021. https://blog.google/threat-analysis-group/updates-about-government-backed-hacking-and-disinformation/.

"Inauthentic Behavior Report." 2020, accessed April 15, 2021. https://about.fb.com/wp-content/uploads/2020/10/Inauthentic-Behavior-Report-October-2020.pdf.

"Information Operations." 2018, accessed April 14, 2021. https://transparency.twitter.com/en/reports/information-operations.html.

Isaac, Mike. 2019. "Facebook Finds New Disinformation Campaigns and Braces for 2020 Torrent." *New York Times,* accessed April 17, 2021. https://www.nytimes.com/2019/10/21/technology/facebook-disinformation-russia-iran.html.

Jeong, Sarah. 2018. *The Internet of Garbage.* Vox Media.

Kahn, Matthew. 2018. "Document: Special Counsel Indicts Russian Nationals and Entities." *Lawfare,* accessed April 14, 2021. https://www.lawfareblog.com/document-special-counsel-indicts-russian-nationals-and-entities.

Kang, Cecilia, Nicholas Fandos, and Mike Isaac. 2017. "Tech executives are contrite about election meddling, but make few promises on capitol hill." *The New York Times [Online], October 31 2017,* accessed April 13, 2021. https://www.nytimes.com/2017/10/31/us/politics/facebook-twitter-google-hearings-congress.html.

Keller, Daphne. 2021. "Some Humility About Transparency." *Stanford Center for Internet and Society,* accessed April 17, 2021. https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.

"The Lawfare Podcast: Alex Stamos on the Hard Tradeoffs of the Internet." 2020. *Lawfare,* accessed February 14, 2020. https://www.lawfareblog.com/lawfare-podcast-alex-stamos-hard-tradeoffs-internet.

Leerssen, Paddy, Jef Ausloos, Brahim Zarouali, Natali Helberger, and Claes H de Vreese. 2018. "Platform Ad archives: Promises and Pitfalls." *Internet Policy Review* 8 (4).

Legum, Judd. 2019. "Facebook Allows Prominent Right-Wing Website to Break the Rules." *Popular Information,* accessed February 8, 2020. https://popular.info/p/facebook-allows-prominent-right-wing.

Lim, Gabrielle. 2020. "The Risks of Exaggerating Foreign Influence Operations and Disinformation." *Centre for International Governance Innovation,* accessed April 15, 2021. https://www.cigionline.org/articles/risks-exaggerating-foreign-influence-operations-and-disinformation.

Lumen. 2019. Accessed October 22, 2021. https://www.lumendatabase.org/pages/about.

Marwick, Alice E, and Danah Boyd. 2011. "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience." *New media & society* 13 (1): 114–133.

McGregor, Shannon. 2020. "What Even Is 'Coordinated Inauthentic Behavior' on Platforms?" *Wired,* accessed April 15, 2021. https://www.wired.com/story/what-even-is-coordinated-inauthentic-behavior-on-platforms/.

Newton, Casey. 2020. "Getting Rid of QAnon Won't Be as Easy as Twitter Might Think." *The Verge,* accessed April 15, 2021. https://www.theverge.com/interface/2020/7/23/21334255/twitter-qanon-ban-facebook-policy-enforcement-political-candidates.

Newton, Casey. 2021. "How YouTube Failed the 2020 Election Test," accessed April 15, 2021. https://www.platformer.news/p/how-youtube-failed-the-2020-election.

Nimmo, Ben, et al. 2021. "Spamouflage Breakout: Chinese Spam Network Finally Starts to Gain Some Traction," https://graphika.com/reports/spamouflage-breakout/.

Office of the Director of National Intelligence. 2017. "Assessing Russian Activities and Intentions in Recent US Elections." *Intelligence Community Assessment,* https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Oremus, Will. 2020. "What Pornhub and Peloton Have in Common With Facebook." *Medium,* https://onezero.medium.com/what-pornhub-and-peloton-have-in-common-with-facebook-d46dbef09b55.

"Platform manipulation and spam policy." 2020, https://help.twitter.com/en/rules-and-policies/platform-manipulation.

Pozen, David E. 2020. "Seeing transparency more clearly." *Public Administration Review* 80 (2): 326–331.

Rid, Thomas, and Ben Buchanan. 2015. "Attributing cyber attacks." *Journal of Strategic Studies,* 4–37.

Romm, Tony. 2017. "Facebook Will Help Some Users Figure out If They Saw Russian Propaganda during the 2016 US Presidential Election." *Vox,* https://www.vox.com/2017/11/22/16689744/facebook-google-twitter-russia-election-disclosure.

———. 2020. "Facebook Disables Russian and Iranian Efforts to Manipulate Users Raising New 2020 Election Fears." *Washington Post,* accessed April 17, 2021. https://www.washingtonpost.com/technology/2020/02/12/facebook-russia-iran-myanmar-disinformation/.

Roth, Yoel. 2019. https://twitter.com/yoyoel/status/1092587833020182528.

Ruiz, Rebecca R, and Mark Landler. 2017. "Robert Mueller, former FBI director, is named special counsel for Russia investigation." *The New York Times,* accessed April 14, 2021. https://www.nytimes.com/2017/05/17/us/politics/robert-mueller-special-counsel-russia-investigation.html.

Santa Clara Principles. n.d. "Santa Clara Principles on Transparency and Accountability in Content Moderation," https://santaclaraprinciples.org.

Scott, Mark, and Elisa Braun. 2020. "France Feuds with Facebook over Disinformation Claims." *POLITICO,* accessed May 15, 2021. https://www.politico.eu/article/france-facebook-disinformation/.

Silverman, Craig, et al. 2021. ""I Have Blood On My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation." *BuzzFeed News,* accessed February 20, 2021. https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo.

Singh, Spandana, and Leila Doty. 2021. "The Transparency Report Tracking Tool." *New America,* https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/.

Stamos, Alex. 2017. "An Update On Information Operations On Facebook." *Facebook Newsroom,* https://about.fb.com/news/2017/09/information-operations-update/.

———. 2018a. "Authenticity Matters: The IRA Has No Place on Facebook." *Facebook Newsroom,* https://about.fb.com/news/2018/04/authenticity-matters/.

———. 2018b. "How Much Can Companies Know About Who's Behind Cyber Threats?" *Facebook Newsroom,* https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/.

Stanley-Becker, Isaac. 2020. "Pro-Trump Youth Group Enlists Teens in Secretive Campaign Likened to a 'Troll Farm,' Prompting Rebuke by Facebook and Twitter." *Washington Post,* accessed May 16, 2021. https://www.washingtonpost.com/politics/turning-point-teens-disinformation-trump/2020/09/15/c84091ae-f20a-11ea-b796-2dd09962649c_story.html.

Stubbs, Jack. 2020. "Exclusive: Russian Operation Masqueraded as Right-Wing News Site to Target US Voters—Sources." *Reuters,* https://reut.rs/3iiXOzK.

Suzor, Nicolas P, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation." *International Journal of Communication* 13:18.

Team, The YouTube. 2020. "Managing Harmful Conspiracy Theories on YouTube," https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/.

Tenbarge, Kat. 2020. "YouTube Channels Made Money off of Fake Election Results Livestreams with Thousands of Viewers." *Insider,* accessed April 15, 2021. https://www.insider.com/youtube-fake-election-results-livestreams-monetized-misinformation-2020-11.

"The State of Influence Operations 2017–2020." 2020, https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf.

TikTok. 2021. "Community Guidelines—Integrity and Authenticity," accessed May 16, 2021. https://www.tiktok.com/community-guidelines?lang=en#37.

Timberg, Craig. 2020. "Parler and Gab, Two Conservative Social Media Sites, Keep Alleged Russian Disinformation up, despite Report." *Washington Post,* accessed May 15, 2021. https://www.washingtonpost.com/technology/2020/10/07/russian-trolls-graphika-parler-gab/.

Timberg, Craig, and Shane Harris. 2021. "Chinese Network of Fake Accounts Targets Trump with English-Language Videos." *Washington Post,* accessed May 16, 2021. https://www.washingtonpost.com/technology/2020/08/12/china-video-network-trump/.

Tumblr Help Center. 2018. "Public Record of Usernames Linked to State-Sponsored Disinformation Campaigns," accessed April 14, 2021. https://tumblr.zendesk.com/hc/en-us/articles/360002280214-Public-record-of-usernames-linked-to-state-sponsored-disinformation-campaigns.

Tumblr Staff. 2018. "We're Taking Steps to Protect against Future Interference in Our Political Conversation by State-Sponsored Propaganda Campaigns," accessed April 14, 2021. https://staff.tumblr.com/post/172170432865/were-taking-steps-to-protect-against-future.

Twitter Public Policy. 2017. "Update: Russian Interference in the 2016 US Presidential Election." *Twitter Blog,* accessed April 14, 2021. https://blog.twitter.com/en_us/topics/company/2017/Update-Russian-Interference-in-2016–Election-Bots-and-Misinformation.html.

———. 2018. "Update on Twitter's Review of the 2016 US Election." *Twitter Blog,* accessed April 14, 2021. https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html.

"Coordinated Harmful Activity." n.d. *Twitter Rules,* accessed April 2, 2021. https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity.

Twitter Safety. 2019. "Disclosing New Data to Our Archive of Information Operations." *Twitter Blog,* https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.

———. 2021. "Disclosing Networks of State-Linked Information Operations," accessed April 14, 2021. https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-.html.

"Platform Manipulation." n.d. *Twitter Transparency Center,* accessed April 15, 2021. https://transparency.twitter.com/en/reports/platform-manipulation.html.

U.S. Senate Select Committee on Intelligence. 2018. "New Reports Shed Light on Internet Research Agency's Social Media Tactics," accessed April 14, 2021. https://www.intelligence.senate.gov/press/new-reports-shed-light-internet-research-agency%E2%80%99s-social-media-tactics.

u/KeyserSosa. 2019. "Reddit Security Report – October 30, 2019." *R/Announcements,* accessed April 14, 2021. https://www.reddit.com/r/announcements/comments/dpa8rn/reddit_security_report_october_30_2019/.

u/spez. 2018. "In Response to Recent Reports about the Integrity of Reddit, I'd like to Share Our Thinking." *R/Announcements,* accessed April 14, 2021. https://www.reddit.com/r/announcements/comments/827zqc/in_response_to_recent_reports_about_the_integrity/.

u/worstnerd. 2019. "An Update on Content Manipulation… And an Upcoming Report." *r/redditsecurity,* https://www.reddit.com/r/redditsecurity/comments/d6l41l/an_update_on_content_manipulation_and_an_upcoming/.

Vilmer, Jean-Baptiste Jeangène. 2019. "The "Macron Leaks" Operation: A Post-Mortem," https://www.atlanticcouncil.org/wp-content/uploads/2019/06/The_Macron_Leaks_Operation-A_Post-Mortem.pdf.

Walker, Kent, and Richard Salgado. 2017. "Security and Disinformation in the U.S. 2016 Election," accessed April 14, 2021. https://blog.google/outreach-initiatives/public-policy/security-and-disinformation-us-2016-election/.

Wardle, Claire. 2020. "The Media Has Overcorrected on Foreign Influence." *Lawfare,* accessed February 1, 2021. https://www.lawfareblog.com/media-has-overcorrected-foreign-influence.

Wardle, Claire, and Hossein Derakhshan. 2017. "Information Disorder:Toward an Interdisciplinary Framework for Research and Policy Making," https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Weedon, Jen, et al. 2017. "Information Operations and Facebook," accessed September 7, 2017. https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf.

Wong, Julia Carrie. 2021a. "'Facebook Isn't Interested in Countries like Ours': Azerbaijan Troll Network Returns Months after Ban." *The Guardian,* accessed April 15, 2021. http://www.theguardian.com/technology/2021/apr/13/facebook-azerbaijan-ilham-aliyev.

———. 2021b. "Facebook Knew of Honduran President's Manipulation Campaign – and Let It Continue for 11 Months." *The Guardian,* accessed April 15, 2021. http://www.theguardian.com/technology/2021/apr/13/facebook-honduras-juan-orlando-hernandez-fake-engagement.

———. 2021c. "Facebook Planned to Remove Fake Accounts in India—until It Realized a BJP Politician Was Involved." *The Guardian,* accessed April 15, 2021. http://www.theguardian.com/technology/2021/apr/15/facebook-india-bjp-fake-accounts.

———. 2021d. "How Facebook Let Fake Engagement Distort Global Politics: A Whistleblower's Account." *The Guardian,* accessed April 15, 2021. https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang.

———. 2021e. "Revealed: The Facebook Loophole that Lets World Leaders Deceive and Harass Their Citizens." *The Guardian,* accessed April 15, 2021. http://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation.

Wong, Julia Carrie, and Olivia Solon. 2018. "Facebook Releases Content Moderation Guidelines—Rules Long Kept Secret." *The Guardian,* https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules.

Authors

Camille François is a Doctoral Candidate at the French Institute of Geopolitics at University Paris 8, a lecturer at the Columbia University School of International and Public Affairs (SIPA), and an Affiliate at the Harvard Berkman-Klein Center for Internet & Society.

evelyn douek is a Doctoral Candidate at Harvard Law School, a Senior Research Fellow at the Knight First Amendment Institute at Columbia University, and an Affiliate at the Harvard Berkman Klein Center For Internet & Society.

Acknowledgements

We thank Emma Llanso, Robert Gorwa, Shelby Grossman, David O'Brien, Alicia Wanless, and anonymous reviewers for their thoughtful comments on previous drafts of this article. Collaboration on this work was a little coordinated, but 100% authentic.

In her previous position at Graphika, Camille Francois worked with platforms mentioned throughout this paper (e.g., Facebook, Google, and Pinterest) on issues related to information operations on their services. She is also an author of several reports mentioned throughout this article, including the expert report on the Internet Research Agency provided to the US Senate Select Intelligence Committee in 2017, which relied on data provided by the platforms to the US Senate. Some of the Graphika reports cited in this paper benefited from access to non-public data on CIB campaigns provided by Facebook. She was a Principal Researcher at Google between 2015 and 2018.

Data Availability Statement

Not applicable.

Funding Statement

Not applicable.

Ethical Standards

Not applicable.

Keywords

Information operations; content moderation; social media platforms; transparency reporting; election interference; coordinated inauthentic behavior