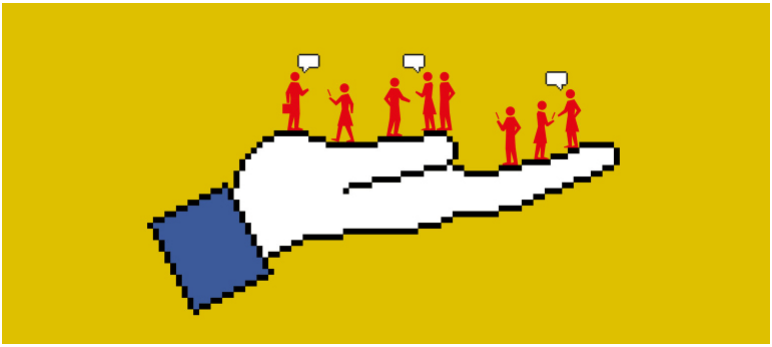
THE TECH GIANTS, MONOPOLY POWER, AND PUBLIC DISCOURSE



The Rise of Content Cartels

By Evelyn Douek

KNIGHT
FIRST AMENDMENT
INSTITUTE
at Columbia University



In November 2019, the Knight First Amendment Institute convened a major symposium at Columbia University, titled "The Tech Giants, Monopoly Power, and Public Discourse," to address concerns arising from the dominance of a small number of technology companies over a wide range of economic and expressive activity. The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, economics, information studies, journalism, political science, and other disciplines, the essays focus on two questions: how and to what extent the technology giants' power is shaping public discourse, and whether anti-monopoly tools might usefully be deployed to expose or counter this power.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; Sarah Guinee, Research Fellow; and Madeline Wood, Communications and Research Coordinator.

The full series is available at knightcolumbia.org/research/

The Rise of Content Cartels

By Evelyn Douek

INTRODUCTION 5

CONTENT CARTEL CREEP 7

CARTELS AS CUSTODIANS 15

THE COSTS OF CONTENT CARTELS 23

GUARDRAILS AND GOVERNANCE 32

CONCLUSION 38

NOTES 39

INTRODUCTION

THE FEAR THAT A SINGLE ACTOR can decide what can or cannot be said in large parts of the online public sphere has led to growing calls for measures to promote competition between digital platforms. At the same time, others are demanding greater cooperation between the custodians of the public sphere. These pressures are not necessarily at odds, but some work needs to be done to reconcile them. To what extent should platforms have consistent content moderation policies? If standards and guardrails are imposed on the public sphere, should platforms work together to ensure that the online ecosystem as a whole realizes these standards, or would society benefit more if it is every platform for itself?

These are questions that need answering as society and regulators demand that platforms become more responsible gatekeepers. Economic competition alone cannot solve many of the speech-related pathologies in online discourse. Collaboration between tech platforms on especially intractable problems allows us to break free of the false dichotomy between too few online gatekeepers holding too much power, on the one

hand, and a fragmented online public sphere constituted by multiple fiercely competitive platforms, on the other.

But this collaboration must be done in a way that keeps a broader sense of the public interest at its core. Currently, most collaboration takes the form of rushed ad hoc alliances of convenience that arise in response to particular crises. The pressure to *do something* can lead to the creation of systems and structures that serve the interests of the very tech platforms that they seek to rein in. I call these *content cartels*: arrangements between platforms to work together to remove content or actors from their services without adequate oversight. These come in various guises; they can be demanded, encouraged, participated in, or unheeded by regulators. But they share the characteristic that they compound the existing lack of accountability in platform content moderation.

In this paper, I begin by tracing the origin and spread of content cartels in Part I, showing that content cartels are the proposed response to an increasing number of pathologies in online discourse. Part II examines the impulses behind demands for greater cooperation and the ways in which such cooperation can be beneficial. Part III explores the failures of the current arrangements and the threats they pose to free speech. Part IV sets an agenda for developing the tools to create productive and legitimate cooperation between platforms in those areas where it can be beneficial or has become inevitable.

This paper has two goals. The first is to raise the alarm about a possible future coming into view, of unaccountable content cartels making decisions about the parameters of online discourse in a way that is just as problematic as an unaccountable monopoly. The second is to explore what can be beneficial about collaborative efforts and what might redeem them. This is a pivotal moment in the management of public discourse, and the structures built now should serve enduring values. We need not settle for institutions that stick band-aids on some problems but do not serve the deeper goal of building trust in online speech governance.

CONTENT CARTEL CREEP

THE PAST FEW YEARS have seen increasing demands for platforms to collaborate in fending off certain perceived threats created by online speech. These don't resemble traditional cartels: They are not hidden but touted, and they are widely seen as beneficial or even necessary. They are also fragmented; rather than a single cartel-like agreement between set members, these arrangements are taking place in different ways in different spheres. But from small beginnings, content cartels are enveloping ever more difficult and contentious areas of online discourse.

The Pilot

Despite the legendarily ruthless competitiveness of Silicon Valley, platforms have long cooperated on the removal of child sexual abuse material (CSAM). Researcher Hany Farid and Microsoft developed a digital fingerprinting technology called photoDNA in 2009.[1] This technology allows images to be "hashed" and added to databases run by nonprofits such as the National Center for Missing and Exploited Children (NCMEC)

and the U.K.-based Internet Watch Foundation (IWF). Companies can then use these databases to automatically and preemptively prevent copies of included images from being uploaded to their services. Facebook deployed photoDNA in 2010, Twitter in 2011, and Google in 2016. "In addition to these titans of technology, photoDNA is now in worldwide deployment," writes Farid.[2] Overall, this has been seen as a success story for tech collaboration.[3]

Even in the case of CSAM, there were early concerns about this model of centralized censorship and its opacity.[4] Despite CSAM being a relatively definable category of content, there can still be unexpected collateral damage without adequate procedural checks. The mistaken addition of the album art for the Scorpions album *Virgin Killer* by the IWF to its Child Abuse Imagery Database in 2008 not only caused ISPs to block access to the band's Wikipedia entry but also prevented most U.K. web users from editing the entire Wikipedia domain, for example.[5] There have also been concerns raised about governance practices and independence.[6] But in general, because the harm of CSAM is so great, the category of content so specific, and the agreement on both of these factors effectively universal, this cooperation between law enforcement, civil society, and the major internet giants is seen as an important institutional innovation to combat the explosion in CSAM caused by the internet.

This is a pattern that will repeat in the story that follows. First, there is the identification of a threat caused by online speech that can be more effectively tackled through coordination. Initial concerns are voiced about the human rights implications of such centralized censorship measures. But ultimately those concerns give way to the view that the seriousness of the threat justifies such measures, provided that the domain is kept carefully circumscribed.

Spread to Terrorist Content

Specifically invoking the success of photoDNA, similar technology was developed to identify terrorist and extremist content. The model was the same: a database of content with hashing technology to check uploads to platform services to prevent the posting of infringing content. But

initially, in 2016, social media companies were hesitant. Executives expressed concern that, because the definition of "terrorist speech" is far less certain and far more contested than CSAM, content that ought to be protected from censorship might find its way into the joint database.[7] This concern is especially acute with respect to countries where the label "terrorist" is used to silence dissent and opposition. If a government or an industry partner added an image to the database, this would facilitate the image's removal from a range of platforms and services, making it an intentionally nimble and powerful tool to meet the growing threat of online extremist content. But this power comes with risks. As one tech officer put it, "As soon as governments heard there was a central database of terrorism images, they would all come knocking."[8]

And yet, by the end of 2016, platforms had changed their tune and announced plans for the database. A mere three years later, the database—run by the industry-led Global Internet Forum to Counter Terrorism (GIFCT)—is a flagship initiative for disrupting terrorists' use of platforms. Facebook and Google representatives appearing before a Senate committee held up the GIFCT as a sign of their more responsible custodianship of the public sphere and touted expansion plans.[9] Monika Bickert of Facebook told the senators that "one of the things that we've done over time is expand the mandate of the [GIFCT] … . [W]e are sharing a broader variety of violent incidents."[10] The number of participants is also expanding.[11] Facebook recently open-sourced the technology it uses to match hashed videos against databases so that industry partners, smaller developers, and nonprofits can also use them to identify abusive content and share hashes.[12] What changed?

This rapid turnaround followed a reversal in public sentiment toward tech and increased government pressure. European Union legislators demanded action after terrorist attacks in Paris and Brussels in late 2015, threatening legislation if companies did not adopt "voluntary" initiatives.[13] This led directly to the establishment of the GIFCT.[14] The livestreaming and viral spread of the Christchurch massacre in March 2019 was a similarly pivotal moment for the initiative. In the wake of that tragedy, public and governmental demands for greater action from tech companies has resulted in (so far) voluntary commitments to expand

the GIFCT and its resources.[15] This increased capacity of the GIFCT was credited with preventing a livestreamed shooting modeled on the Christchurch attack in Halle, Germany, from going viral.[16] Governments have since announced plans to ramp up task forces dedicated to working with the database, again specifically invoking the CSAM model.[17]

This much of the story treads familiar ground. Danielle Citron has memorably coined this evolution "censorship creep" to describe how a tool intended to help execute one type of speech regulation (the removal of CSAM) comes to be used for others (such as terrorist content).[18] But my focus is on a different aspect: It is not only the areas of censorship that have expanded but also companies' cooperation. In another timeline, companies could instead have engaged in an arms race to develop the best technology to remove such content and find favor from regulators and users. Instead, they are agreeing to cooperate on the enforcement of more uniform norms of what should be available online. Although voluntary, this collaboration does not necessarily arise out of a sense of moral obligation—beyond regulatory threats, there are also reputational reasons it could be advantageous for companies to present a united front and avoid being singled out for difficult choices in areas not considered core to their product.[19] As discussed below, there are ways this can be socially beneficial and not only serve the interests of the platforms involved: The premise of these institutional innovations is that platforms working together are more readily able to meet threats posed by the spread of disfavored content in the public sphere.

Again, however, the dynamic was the same: an unlikely collaboration between highly competitive firms, at first resisted and carefully circumscribed but ultimately embraced and expanded, in a form that lessens public pressure without increasing public accountability.

"Coordinated Inauthentic Behavior"

Content cartel creep is now spreading to the fight against foreign influence campaigns. This is an area of increased focus and activity, both from perpetrators and defenders alike, in a manner reminiscent of the increased concern about terrorist content a few years before the establishment of the GIFCT.

Online influence campaigns are often sophisticated, cross-platform operations. According to a report prepared for the Senate Intelligence Committee, during the 2016 election Russia's Internet Research Agency "operated like a digital marketing agency: develop a brand … build presences on all channels *across the entire social ecosystem*, and grow an audience with paid ads as well as partnerships, influencers, and link-sharing. They created media mirages: *interlinked information ecosystems* designed to immerse and surround targeted audiences."[20] This mirrors earlier experience with terrorists' use of the internet, where groups used different platforms for different purposes, leveraging their unique affordances.[21] This has only grown more true of recent influence campaigns. A Digital Forensic Research Lab investigation of a suspected Russian intelligence operation in 2019 found that it spanned over 30 platforms in nine languages.[22]

In detecting these types of influence campaigns, tech companies increasingly rely on sharing information with each other and governments.[23] This kind of collaboration may be indispensable to counter-influence efforts.[24] As a result, calls for collaboration are getting louder, and sharing and coordination has resulted in multiple simultaneous takedowns of accounts and content linked to related influence campaigns that operate across different platforms.[25]

The case for cooperation in this context is simple and intuitive: If information operations work across different platforms, so too should the efforts to combat them. Researchers express dismay when pages and groups blocked on Facebook continue to operate accounts on Twitter, YouTube, and Instagram.[26] Even with information sharing, it may be months before smaller platforms identify accounts related to influence campaigns taken down by other platforms.[27]

Platform information sharing in this context is currently informal and ostensibly ad hoc. It is also opaque: In September 2019, in a secretive gathering from which few details emerged, U.S. law enforcement representatives met with Facebook, Google, Microsoft, and Twitter to discuss election interference ahead of the 2020 election.[28] But there are calls for such collaboration to expand and become systematic. The bipartisan Senate Intelligence Committee report on Russia's use of social media in

the 2016 election again specifically invoked the NCMEC model in calling for Congress to "consider ways to facilitate productive coordination and cooperation between U.S. social media companies and the pertinent government agencies."[29] Rep. Ro Khanna has floated legislation to encourage such cooperation.[30] Likening the model to banks working together to share data to detect fraud, Khanna wants a consortium for sharing disinformation threats to ensure that a bad actor can't be "kicked off Facebook and just go open an account on Twitter."[31] Similarly, a Stanford report has called for building a body to coordinate efforts to combat disinformation, modeled on successful examples like the nonprofit Financial Services Information Sharing and Analysis Center, which works to ensure resilience and continuity of global financial services and has its own budget, staff with security clearances to receive threat briefings, technical tools, and the power to facilitate sharing and consensus building among the membership.[32] It suggested that Congress should create specific antitrust safe harbors to allow such coordination around the sharing of information and banning of proscribed content.[33] An NYU report also called for the formation of a "permanent intercompany task force devoted to fighting disinformation," specifically invoking platform collaboration through the GIFCT and photoDNA as a model of the "spirit of cooperation that ought to infuse the push to limit disinformation."[34]

The Stanford report highlights an essential caveat that is absent from current collaborations: Tighter coordination needs to have third-party oversight to be credible and legitimate.[35] Without oversight, there is little to ensure that individual rights will be adequately respected. Such oversight is often part of the original vision of these collaborations, at the point of the conversation when free speech concerns are given their fullest hearing. But, as discussed further below, so far this has not been realized in practice.

The Creep Continues

Similar themes permeate the conversation about the threat of synthetic media, more commonly referred to as "deepfakes." As concerns rise about the potential for deepfake videos to swing elections or ruin lives through an invasion of sexual privacy, there are calls to create a GIFCT-style

database of such media to prevent their spread online.[36] Even further from its roots in the CSAM database, the societal conversation about how to define the category of synthetic media that should be removed from social media has barely begun. While CSAM material is harmful in every form and context, this cannot be said of synthetic media, which can be used for educational or entertainment purposes.

Further still along this path are calls for a more collaborative, whole-of-internet approach to be adopted toward hate speech and extremism more broadly. Rep. Khanna, for example, has suggested that his proposed consortium could facilitate information-sharing not only with respect to disinformation but also with information about hate speech and other inflammatory content.[37] A U.K. white paper on "online harms" proposed a new regulator to foster a "culture of cooperation" between companies, noting that "[u]sers perpetrating harm often move between platforms, especially to behave illegally and disseminate illegal content. A greater level of cooperation between platforms by sharing observations and best practices to prevent harms spreading from one provider to another will be essential."[38] A candidate for the U.S. Democratic Party nomination for president called for uniformity across the industry, saying, "[Y]ou can't have one set of standards for Facebook and another for Twitter."[39] Again, many of these proposals are framed as natural progressions of the previous models. As former Facebook Chief Security Officer Alex Stamos told Congress last year, "[The GIFCT] has been somewhat successful in building capabilities in smaller members while creating a forum for collaboration among the larger members. It is time to follow this initial foray with a *much more ambitious coordinating body* between tech companies focused on adversarial use of their technologies."[40]

Other forms of platform alignment could occur through speech decisions being repackaged as trust and safety tools and services that platforms can offer each other. Jigsaw, a technology incubator created by Google, has developed Perspective API, which helps platforms moderate content by using machine learning models to detect "toxic" comments.[41] It is used by Latin America's second-largest social platform, Taringa!, for example.[42] This creates the danger that biases in one platform's data will find their way into how other platforms moderate. In the case of Perspec-

tive, researchers have raised concerns that the tool has a disparate impact on already marginalized communities.[43] Another example is Microsoft's announcement[44] in January that it is developing software to recognize online predators attempting to lure children for sexual purposes, which it will release to other companies free of charge.[45]

Stamos has suggested that smaller platforms' need for assistance in content moderation may go beyond the supply of mere tools and might even create its own industry:

> Say you start a company and all of a sudden the Nazis take over … there are very few people you can call to come help you with your Nazi problem, or with your child safety problem or your suicide problem … you can't expect that every company builds all these things from scratch … . You could see Facebook turning their experience scaling trust and safety into an actual service they provide to all of the smaller companies who can't afford to hire thousands of people themselves.[46]

All such proposals are a long way from the carefully circumscribed CSAM databases. But if you build it, they will come. And often, once there is a decision that certain content cannot be left to the marketplace of ideas, the decision quickly follows that platforms should cooperate to ensure consistent enforcement.

CARTELS AS CUSTODIANS

WHAT EXPLAINS THE PHENOMENON of content cartel creep? There are a number of forces at work that serve as rationalizations for the need for platforms to cooperate more and across a greater variety of contexts.

Competition Is Not a Cure-All

Content cartel creep is at least in part an implicit acknowledgment that the relationship between economic competition in the platform market and a healthy public sphere is complicated. There are a number of speech-related pathologies in the current online environment that competition alone cannot solve. Of course, monopoly power over public discourse is a significant concern—it is not only government officials who should not be able to declare what is "orthodox in politics, nationalism, religion, or other matters of opinion."[47] But while greater competition will alleviate this threat (and, importantly, other non-speech harms), there are a number of speech-related harms for which competition is not a sufficient answer.[48] Indeed, the call for platforms to be better gate-

keepers and "custodians of the internet," as Tarleton Gillespie has so evocatively labeled them,[49] can at times be in tension with concerns that platforms hold too much power to determine the bounds of acceptable discourse. Jillian C. York and Ethan Zuckerman observe that decentralization as a cure for the concentration of power in the major platforms "replace[s] one set of moderation problems—the massive power of the platform owner—with another problem: the inability to remove offensive or illegal content from the [i]nternet."[50] Aviv Ovadya calls this the "'magical decentralization fallacy'—the mistaken belief that decentralization on its own can address *governance* problems."[51] Ovadya argues that in cases of misinformation and harassment, "decentralization just turns a hard centralized problem into a harder coordination problem."[52] European Union Competition Commissioner Margrethe Vestager has more colorfully likened the problem to the hydra: chop off one head and "two or seven [come] up—so there is a risk you do not solve the problem[;] you just have many more problems."[53]

These critiques highlight that focusing on the threat of monopoly power alone risks overlooking the many other concerns that exist in the public sphere as it is currently constituted. Fears about the spread of hate speech, disinformation, election interference, filter bubbles and echo chambers, facilitation of government censorship, and the coarsening of public discourse are all on the rise, and none would necessarily disappear if large platforms did.[54] As momentum for antitrust action against big tech grows, so too does the recognition that there are many problems antitrust cannot solve.[55] Some pathologies, such as echo chambers[56] or the spread of low-quality or radicalizing content driven by optimization for engagement,[57] could be exacerbated in a hypercompetitive environment. In other cases, potential structural fixes, such as increased friction to slow down the spread of harmful content or increase information fidelity and cognitive autonomy,[58] will not be supplied by the market alone. User preference often seems to manifest as friction for thee, but not for me. Therefore, as Commissioner Vestager recently said, "[I]t's understandable that people sometimes think of competition as a panacea, a universal answer to all society's problems. But it can't be that. … [I]f, as a society, we want to lay down fundamental standards … then what we

need is not more competition enforcement. We need regulation."[59]

This is underscored by the fact that certain online speech-related harms are not confined to one platform or business model.[60] Time and again, the same harms recreate themselves on different platforms. Telegram is a popular encrypted nonprofit messaging app that does not algorithmically rank user content. Even this plain-vanilla app can be employed for different ends: It has been both a tool for activists in Hong Kong and Iran as well as a meeting place for far-right extremists.[61] When Telegram purged a number of ISIS accounts, they migrated to a platform called Riot, and when Riot also cracked down the accounts found refuge in TamTam, a platform very similar to Telegram but smaller and with limited capacity to respond.[62] WhatsApp, another non-algorithmically mediated messaging service, has been implicated in the spread of misinformation, hate speech, and violence in places as diverse as India, Brazil, and Nigeria.[63] These concerns about WhatsApp are now echoing in the fears about content on TikTok arising as that platform becomes more popular.[64] This illustrates that these problems permeate across internet platforms. As Alexis Madrigal put it, "If you were to declare that both WhatsApp and Facebook were problems, you come uncomfortably close to admitting that mobile communications pose fundamental challenges to societies across the world. Which … there is a decent case for."[65] Because, while the problem with Facebook is of course, in some part, Facebook,[66] "part of the problem with Facebook and other platforms is *people*, easily distracted, highly susceptible to misinformation, and prone to herd behavior."[67]

Standards as Barriers to Entry

If competition cannot be relied upon to solve current online speech-related pathologies, then standards can be imposed through regulation or social pressure. But without collaboration, such standard-setting can itself create barriers to entry and harm competition. While larger platforms have developed tools and can devote the resources to the problems that inevitably accompany providing a platform for user-generated content, this is not true for smaller companies. Commissioner Vestager notes that requiring consistent responsibilities from big and small companies

alike risks giving many potential disruptors "no chance of competing."[68]

The problem is especially acute for small platforms and startups, which may find themselves the host of certain kinds of content that they didn't anticipate when starting, for example, a gaming livestreaming platform or a community for knitters and crocheters. April Glaser explored the example of Discord, a chat platform for gamers that found itself host to white supremacist groups: "How a company like Discord should deal with this activity isn't necessarily obvious. Discord doesn't have the resources of Facebook or Twitter, which have drawn a clearer line of what kinds of speech and activity they tolerate on their platforms, and Discord most likely can't dedicate large teams to building machine learning tools aimed at ferreting out hate."[69] As an indicative figure, YouTube spent over $100 million on developing the technology it uses to identify copyright violations[70]—a totally unimaginable amount for smaller players. Facebook's head of counterterrorism puts it bluntly: Plenty of startups are "just trying to keep the lights on, and nobody realizes that they need to be dealing with this."[71] Cloudflare—a cloud service provider—explained, when announcing its decision to share CSAM tools to all its clients for free, that "as the regulatory hurdles around dealing with incredibly difficult issues like CSAM continue to increase, many of them lack access to sophisticated tools to scan proactively for CSAM. You have to get big to get into the club that gives you access to these tools, and, concerningly, being in the club is increasingly a prerequisite to getting big."[72]

Collaboration is one way of resolving this tension. David Kaye, the U.N. special rapporteur for freedom of expression, for example, has called for this in the context of developing tools to detect hate speech, saying that "[t]he largest companies should bear the burden of these resources and share their knowledge and tools widely, as open source, to ensure that smaller companies, and smaller markets, have access to such technology."[73]

Cross-Platform Threats

Another key driver of cartel creep is the fact that certain threats may not be identified, let alone addressed, without companies working together. This is especially true in the context of disinformation campaigns or ter-

rorist groups which, as described above, often operate across a multitude of platforms; in such cases, industry information sharing can be essential to their detection, and lack of coordination can mean that bad actors booted from one site can continue operating on other platforms unencumbered.[74] Social media manipulation investigator Camille François has referenced "a key concern with regard to the current industry responses to viral deception: while disinformation actors exploit the whole information ecosystem in campaigns that leverage different products and platforms, technology companies' responses are mostly siloed within individual platforms (if not siloed by individual products!)."[75] Stamos has similarly argued that "[t]he long tail of social platforms will struggle with information operations unless there are mechanisms for the smaller companies to benefit from the research the large companies can afford. There is some precedent in child exploitation and terrorism that can be built upon."[76]

Cross-platform coordination addresses other problems too. For example, in the cases of extremist content, smaller platforms see an increase in such content on their sites as major platforms crack down.[77] So without collaboration, these problems are not solved, only moved. Smaller platforms that do try to make changes can also find their efforts stymied by the lack of coordination by more mainstream sites. For example, as Pinterest took action against anti-vaccine misinformation, it found its efforts partially frustrated by the failure of other platforms to block such content, allowing users on Pinterest to simply link to it.[78]

Protecting Speech

Importantly, but perhaps counterintuitively, there are ways in which collaboration can result in *greater* speech protection through the development and dissemination of more capable tools. Developing artificial intelligence tools for content moderation at scale is hard and resource-intensive. In some cases, it will not be possible without access to large datasets of the kind that only the biggest platforms have access to. Troublingly, and I return to this below, this means that the struggle for smaller companies and the drive to cartelization is especially strong when content moderation is hardest and most controversial. But in the absence

of tool-sharing, some platforms may resort to blunter measures.

An illustrative example is JustPaste.it, a free content-sharing platform run by one person, a 26 year-old Polish man named Mariusz Żurawek. JustPaste.it unwittingly became a favored tool for those spreading ISIS propaganda.[79] Mariusz struggled with how to address complaints about his platform and how to moderate content in Arabic.[80] Without assistance, one option was to remove Arabic content altogether. By joining the GIFCT, JustPaste.it was able to find a less drastic solution for identifying and removing terrorist propaganda. This is just one illustration of a recurrent dynamic where large platforms will have the nimblest and most nuanced detection tools.

Another way platform cooperation could protect speech is by giving companies political cover to resist governmental pressure to remove speech. Because such "jawboning" is often done away from the public eye, this is hard to evaluate, but for small platforms who might be especially susceptible to such pressure, pointing to an established process for detecting harmful content could help legitimate resistance.

Cleaning Up Your Platform, Not Just Washing Your Hands

An important underlying theme of these developments is the need to think more holistically about the online content ecosystem, rather than taking a stovepipe, platform-by-platform approach to speech-related harms. The implication is often that the true Good Samaritans do not stop at cleaning up their own platform but also think about how their rule enforcement choices will affect the broader public discourse. Otherwise, unilateral action by a platform may solve *their* problem, but not society's. As Oren Segal, director of the Anti-Defamation League's Center on Extremism, said in the context of extremism, "We're having the same conversation every few months … . Instead of talking about 8chan, today it's Telegram. Instead of a Facebook stream, we're talking about Twitch. I think that underscores a broader issue that extremists will migrate from platform to platform."[81]

Content cartels will never be all-encompassing. As one platform employee memorably told Sheera Frenkel of the *New York Times*, refer-

ring to the CEOs of Facebook, Google, and Twitter, respectively, "I could see Mark [Zuckerberg], Sundar [Pichai] and Jack [Dorsey] tripping acid at Burning Man faster than I can see them all getting to a consensus on content policy."[82] But as public pressure and controversy around content decisions increases and becomes increasingly costly, the pressure not to compete on things not core to platforms' business models may also increase. Content moderation is hard, and presenting a united front can prevent any individual company from receiving too much opprobrium for any particular decision. This may be especially true for smaller platforms, which can draft in the wake of higher-profile companies.

As with any cartel, this could create a prisoner's dilemma. An example is the viral spread of a manipulated video of Speaker Nancy Pelosi that falsely made her appear drunk: When Facebook received sustained public criticism after deciding not to take the video down, others, most prominently YouTube, quickly reacted and took the video down, offering reasons of varying degrees of persuasiveness but receiving public credit anyway.[83]

In particularly fraught cases, however, the risk-averse option will often be to act together. An especially notable example of this is the "de-platforming" of high-profile conspiracy theorist Alex Jones. Jones was protected by platforms' desire not to be the arbiters of truth—until he wasn't, all at once.[84] When Apple removed Jones's content from iTunes, other platforms fell like dominoes. Mark Zuckerberg decided to kick him off Facebook at 3 a.m. in a hotel room after hearing about Apple's move.[85] It is unlikely that platforms across the industry arrived at the same conclusion at the same time after a delicate balancing of liberty and dignitary interests by coincidence. The same dynamic can be seen in operation in areas as diverse as platforms announcing rules against nonconsensual pornography in quick succession[86] and the variety of hosts scrambling to prevent 8chan sitting on their servers when Cloudflare made the decision not to host it.[87] These public pressures might be what has led Mark Zuckerberg to hope that Facebook's new oversight board for content decisions will become an industry-wide body[88]—industry consistency can dilute public criticism. But it is not only companies that see benefits in this kind of industry consistency: The nonprofit Article 19, for example, is working

on a proposal for a multi-stakeholder "Social Media Council" accountability mechanism that incorporates platforms across the industry.[89] There is force behind the argument that standards established through legitimate processes should operate consistently across platforms.

It is important to note that forces toward the cartelization of content decisions remain whether or not antitrust action is taken against the major social media platforms to break up individual companies, and whether or not smaller platforms come to displace the current monoliths that dominate the public sphere. These questions are not going away. Content cartels are fundamentally a response to the growing consensus that there are certain areas that need to be placed beyond competition, both in the economic marketplace and the marketplace of ideas. Civil society, users, and lawmakers are demanding more comprehensive responses from social media platforms, and platforms are seeing fewer upsides to resisting these calls in certain areas. Whatever regulatory action is taken with respect to the current tech giants to preserve the vitality of the public sphere, it needs to be done in a way that ensures that the problems of cartelization are not exacerbated.

THE COSTS OF CONTENT CARTELS

ONE WAY TO READ THE STORY of content cartels is as a tale of progress. For some harms created by online speech, collaboration between tech platforms can significantly limit the damage, which is why such collaborations are being pushed. But just as monopoly power over public discourse can be pernicious even when exercised for ostensibly beneficial ends, an opaque cartel may be no better. This part discusses four key ways informal and unregulated cartels threaten to exacerbate underlying problems in current content moderation practices.

Compounding Accountability Deficits

First, content cartels only compound the lack of transparency, due process, and accountability that come about when individual platforms make decisions about what speech should or should not be allowed on the internet. Content moderation more generally is going through a crisis of legitimacy. There is growing awareness about the arbitrary and unaccountable way that tech companies develop and enforce their rules of what is allowed on their platforms. Content cartels exacerbate this—when

platforms act in concert, the actual source of any decision is harder to identify and hold to account.

The GIFCT database is a good example. Emma Llansó has compellingly warned about the database's longstanding transparency and accountability deficits.[90] Nobody apart from consortium members knows what is in the database or who added any piece of content, and there are no established independent mechanisms to audit or challenge inclusions. By creating another layer between affected users and the source of the decision, access to remedy is more attenuated in cases of error even as the effects are magnified. If an image or URL is mistakenly classified as terrorist content by one platform, it is significantly more likely to be removed across all participating platforms instead of the mistake being confined to one service.[91]

As content cartels move into areas where categories are more contested and ambiguous, the likelihood of mistakes rises. As Special Rapporteur Kaye has noted about the GIFCT, "This is not like child sexual abuse, for which there is a consensus around imagery that clearly and objectively meets a concrete definition. Rather, it is asking companies to make legal decisions, and fine ones at that, about what constitutes the elements of terrorism, of incitement to terrorism, of the glorification of terrorism."[92] Companies know this. As Brian Fishman, Facebook's head of counterterrorism, has noted, "[T]errorist content might be shared for legitimate reasons by academics, activists decrying extremism, or journalists. The notion that there are legitimate reasons to share terrorist propaganda significantly distinguishes this kind of content from other types of harmful content found online, most notably child pornography."[93]

Mistakes are far from uncommon. The Electronic Frontier Foundation (EFF) has documented many instances where materials have been removed from social media platforms in error.[94] Famously, the Syrian Archive, a civil society organization that works to preserve evidence of human abuses in Syria, is in an ongoing struggle with YouTube, which has deleted over a hundred thousand of its videos.[95] Companies are not the only ones to make mistakes: In April 2019, the French Internet Referral Unit sent a notice to the Internet Archive requiring the removal of over five hundred URLs it identified as "terrorist propaganda," which included

such problematic content as the entire Project Gutenberg page of public domain texts, a massive Grateful Dead collection, and a page of C-SPAN recordings.[96]

When such decisions are removed from public view and filtered through a collective but opaque institution, it is easier for mistakes to be missed, and for members to shirk blame for any that are found, by making it more difficult to identify the source. This is problematic enough in cases of an innocent mistake, but it also creates a more powerful chokepoint for collateral censorship by governments.[97] Cindy Cohn, executive director of the EFF, put it bluntly: "I wouldn't want a central location where censorship decisions get made."[98] The pressures and motives that lead to mistakes in the context of terrorist and extremist content would crop up if the model of the GIFCT were to expand to other areas such as foreign influence operations, deepfakes, or hate speech. As Farid, a key researcher in the development of hashing technology, has himself observed:

> It is important to understand that any technology such as that which we have developed and deployed can be misused. The underlying technology is agnostic as to what it searches for and removes. When deploying photoDNA and eGlyph, we have been exceedingly cautious to control its distribution.[99]

Because of this potential for abuse, building in third-party oversight and accountability mechanisms from the start is essential. The GIFCT example shows that when institutions are set up as reactions to particular crises, the institutional design may not serve longer-term or broader interests.

Creating a False Patina of Legitimacy

Second, content cartels create the risk of unaccountable errors even as they give decisions a greater patina of legitimacy. Announcements of collaboration create a mirage of progress, without necessarily furthering the resolution of underlying issues. This echoes "greenwashing" in the environmental space or "bluewashing" charges leveled at some United

Nations partnerships; these terms reference the idea that companies can be seen as more eco-friendly or human-rights-observant merely by participating in certain performative arrangements. Similarly, simply by announcing that they are working together or creating institutional auspices for their actions, platforms can look responsive to demands that they do *something*. Often this process of legitimation occurs while sidelining the civil society oversight that was initially envisaged as part of the plan.[100]

The history of the Global Network Initiative (GNI) is illustrative. The GNI focuses on promoting freedom of expression and privacy in the information and communications technology sector; it was formed in 2008 in the wake of scandals in which both Yahoo! and Google were found to be involved in human rights violations in China. In response to public outcry, the two companies participated in the establishment of the GNI, which facilitates various industry consultations and human rights audits in order to establish companies' compliance with human rights commitments. Nevertheless, years later the Edward Snowden revelations showed that the GNI verification mechanisms had failed to reveal the extent of data collection and sharing by many GNI members with the U.S. government.[101] The EFF, a founding member of the GNI, resigned as a result.[102]

Companies similarly appealed to the legitimacy of the GIFCT in the wake of the Christchurch massacre as evidence of their commitment to fighting the spread of violent footage.[103] But when GIFCT members boasted that they had added over 800 new hashes to the database, there was no way to verify what this meant or whether it was a good marker of success.[104] There was, for example, no way to know if these included legitimate media reports that used snippets of the footage, or completely erroneous content, or what proportion of the variants of footage uploaded the figure represented. These deficiencies repeated themselves in the wake of the Halle livestream, even as the platforms were congratulated for their effective response.[105]

Content cartels also further embed and legitimize standards without proper public contestation. Companies and governments can use technical-sounding terms like "hashing" to describe coordinated censorship of material falling under the ill-defined standard of "terrorist propaganda."

This suggests a quasi-scientific neutrality and verifiability to what is ultimately a human and value-laden choice about what should be included. Similarly, choreographed cross-platform takedowns of "networks" deemed to be engaging in "coordinated inauthentic behavior" suggest a certainty and objectivity to what companies have deemed problematic, but external verification is limited.

Company takedowns of networks originating in China aimed at delegitimizing protests in Hong Kong bear this out. Facebook, Twitter, and Google announced takedowns relatively contemporaneously and said they were relying on information sharing between them. But this united front obscured a number of questions. It soon became clear that a number of innocent accounts had been swept up in Twitter's takedown.[106] As one researcher put it, "There's a lot of chaff in this wheat."[107] The sweep was also under-inclusive, with other researchers finding batches of coordinated accounts promoting disinformation narratives that survived the takedowns.[108] When Facebook said that it conducted an internal investigation into suspect behavior in the region "based on a tip shared by Twitter,"[109] it was avoiding the question of why it wasn't investigating the region already or why it located comparatively fewer accounts. Google simply announced that the accounts it took down were "consistent with recent observations and actions related to China announced by Facebook and Twitter" without releasing any examples or data of what it took down.[110] Somewhat ironically, these statements invoke their own coordinated action as a source of legitimacy for the takedowns of coordinated influence operations.

Transparency about standards can be difficult in these cases, because it could allow bad actors to game the system. But complete opacity or a lack of external validation is also problematic, especially when it facilitates subjective censorship of political speech. As Ben Nimmo put it, "It's almost like wherever you look, you're finding this stuff."[111] Which means that takedowns depend on where you look. And when companies look where they are told by industry partners, this can create uneven enforcement.

The upshot is this: Categories such as synthetic media, influence operations, hate speech, and harassment are far removed from the

well-defined category of CSAM. This speech will often be political, and there are contexts in which it has legitimate purposes, such as reporting or research. Because of this ambiguity and contentiousness, in especially difficult or high-profile cases this may increase companies' inclination to act uniformly to avoid bearing public backlash. That is, it is precisely those areas in which lines are blurry and public contestation is important that cartelization becomes most attractive. And, as noted above, it is also in those cases where technical tools are hardest to develop because the category is harder to define. In such cases, an institutional or other collaborative framework can give companies an aura of legitimacy that individual company decisions would lack. For smaller companies especially, it can provide a shield from scrutiny as they deal with the unexpected yet inevitable consequences of starting an online platform. Through these moves, calls for platforms to work together redirect attention away from resolution of the difficult anterior questions of what they should work together *toward*.

Augmenting the Power of the Powerful

Third, content cartels augment the power of already powerful actors by allowing them to decide standards for smaller players. This traces the historical experience with public international organizations where the faith that international efforts to combat international problems were inherently good gave way to the realization that "powerful states and special interests were, in fact, steering them in favour of their own ends."[112] As Julie Cohen writes, "[I]f a particular hub within a dominant network exercises disproportionate control … then networked organization will amplify that hub's authority to set policy."[113]

Again, take the GIFCT. When major platforms decide that something constitutes "glorification of terrorist acts," a notoriously vague category, they make this decision not only for themselves but for all other smaller GIFCT members. As we have seen, this can also be held up as a virtue. Developing detection technology is difficult and costly and beyond the capacity of many small startups. Expanding the GIFCT's accessibility for smaller platforms is a key part of the Christchurch Call[114] for this reason. Furthermore, as more and more jurisdictions impose onerous regulatory

obligations (many with short deadlines and high liability), operating in these markets could become impossible without help removing illegal material.

But where lines are blurrier, the underlying premise of the marketplace of ideas suggests that there should be room for contestation and debate. Cartels do not allow for this. For many smaller companies, it will be a take-it-or-leave-it situation. While participation in the GIFCT does not require automatic removal of any content that matches with the hashing database, the time and resources needed to check individual content will be beyond the capacity of most smaller players. A case study of JustPaste.it's experience joining the GIFCT showed how time-consuming the process was, even with support and reliance on the classifications already in the database.[115] As Brian Fishman, Facebook's counterterrorism lead, put it, when it comes to whether or how to use the GIFCT database, "smaller companies have to make difficult decisions about where to apply limited engineering resources."[116]

This take-it-or-leave-it proposition will hold true for smaller governments too. Larger or more powerful states are imposing obligations that tech platforms often, for ease (or, increasingly, due to legal mandates), apply worldwide. These powerful governments can in some sense become part of the cartels—defining what is permissible material for other countries who did not have a seat at the table. Facebook, for example, accedes to U.S. government "terrorist" designations and removes them from their platforms, even when the U.S. designates other state actors.[117] This also creates enforcement gaps, where platforms enforce against those organizations that governments keep track of but not others. For example, Facebook and Twitter de-platform Islamic State extremists more often than white supremacists in the U.S. because the government does not keep equivalent lists. In this way, these content cartels augment the power of major platforms and governments by facilitating the application of their decisions globally and across the industry.[118]

Furthermore, within multi-stakeholder governance institutions, powerful players can exploit ambiguity and bend the rules to their advantage. As Cohen shrewdly observes, within networked governance arrangements, "power interprets regulatory resistance as damage and routes around

it."[119] Absent the typical limits that are imposed on institutions within a domestic setting, these typically international efforts allow powerful actors to engage in regulatory arbitrage without any institutional check.[120] For example, while Fionnuala D. Ní Aoláin, the U.N. Special Rapporteur on human rights and counterterrorism, has expressed concerns about Facebook's definition of terrorism,[121] there is no mechanism to enforce the international human rights issues she raises against Facebook and so Facebook can accede to the demands of governments who possess the greatest leverage over it.

Suggesting a False Equilibrium in the Marketplace of Ideas

Fourth, homogenization across different platforms, if not made recognizable and overt, can be especially problematic when it comes to speech because it suggests a false consensus about where lines should be drawn. Cartels can create the misleading impression that the end result, i.e., a lack of diversity across platforms, was independently arrived at in multiple environments instead of as a result of a single decision being applied across all of the platforms. Particularly in difficult or controversial areas, it could become increasingly attractive for platforms to adopt a kind of default position instead of facilitating the kind of diversity that could help continually reevaluate the bounds of accepted expression. In some areas such continual reevaluation is not necessary: For example, the desirability and definition of CSAM is quite properly well settled. But, as has been a continual theme of this paper, as pressure grows to expand this model to other areas, the risks are greater because few categories are so determinate.

* * *

A complaint about monopoly power over public discourse is, at heart, a complaint about the use of unaccountable power to demarcate the public sphere in ways that do not adequately serve public ends. This can happen in more ways than one, and content cartels recreate this risk in another form. Mary Anne Franks has powerfully written that to "truly believe

in the 'marketplace of ideas' means to reject speech monopolies and speech cartels, to challenge the hoarding of expressive rights by the most privileged members of society."[122] Current content cartels, as well as the future ones we are likely hurtling toward, allow participants to launder difficult decisions through opaque processes to make them appear more legitimate than they really are and do not mitigate the threat of a handful of actors holding too much power over the public sphere.

GUARDRAILS AND GOVERNANCE

THOSE CONCERNED WITH MONOPOLY POWER over public discourse should similarly be concerned about the rise of content cartels. But is it possible to keep the baby of helpful collaboration and throw out the bathwater of harmful cartels? In some areas and for some problems, platforms working together can be beneficial. But *in which areas* and *how* platforms collaborate is as important as that they do.

When Should Platforms Cooperate?

The scoping question of when to collaborate is fundamental and is currently being answered by the interests of the powerful, in reaction to particular crises. Platforms themselves decide (determined of course in large part by public and regulatory pressure) when, how, and with whom collaboration will occur. An observation from another context is apt: "Reasonable minds may differ as to what the ideal balance between cooperation and resistance might be, but it seems unlikely that this balance should be left to the judgment of a private corporation."[123]

The appropriate level of cooperation is not an easy question. It might

be tempting to say that cartels are always too great a threat to diversity in the marketplace of ideas, but such a response exacts a large cost. If you accept that there need to be standards for speech online,[124] it is difficult to defend the proposition that these standards should not be enforced effectively. In cases where a lack of coordination means simply moving the problem around or, worse, smaller platforms either using blunter tools or not moderating at all, collaboration could be a boon. Should smaller platforms be denied the technology to remove violent propaganda on their platforms (which in many cases ends up there only after being banned by major platforms) in the name of marketplace diversity? Should these platforms be forced to choose between less nuanced hate speech detection tools or a species of free speech absolutism to avoid cartelization? What if this leads to exactly the kind of echo chambers that are most concerning? Homogenization of the public sphere cannot be our only concern.

Another response might be to say that cartels should only exist for categories of content that are well defined and do not implicate value judgments, but this distinction quickly collapses. None of these lines avoid value judgments. Even the most neutral-sounding, such as "coordinated inauthentic behavior," rely on fundamental judgments about required levels of "authenticity"[125] and permissible forms of freedom of association.[126]

Ultimately, the answer will differ in each context. There are some easy cases: When a category of content is illegal and comparatively definable and the effective detection and removal of its presence online depends on technological tools that it is impossible or impractical for all platforms to develop individually, collaboration can be useful—*provided there is appropriate oversight*. CSAM is the obvious case. Another context might be footage of extreme violence like the Christchurch massacre (provided that collaboration is done in a way that prevents creep into less extreme but still controversial content). Facebook has, for example, announced a partnership with law enforcement to use bodycam footage from training exercises to train an AI tool to identify and remove violent first-person footage.[127] Smaller platforms could benefit from the use of such technology but cannot develop it themselves.

At the other end of the spectrum are easy cases too: In matters of platform policy only—where the question is not one of sharing technical tools—and where the category of content is legal and contestable, platforms should have to make and justify their rules openly and on their own. What to do with Alex Jones types, or "glorification of terrorism," or dispiriting but legal posts by public figures should not be decided behind an unaccountable "consensus." Platforms should be forced to transparently justify their individual decisions in accordance with their public rules.

Between these two ends of the spectrum are the hard cases. Ultimately, the answer should depend on an empirical inquiry into factors such as the prevalence of that category of content; the accuracy of the relevant technology; the cost and practicality of small platforms developing similar tools; the relevant risk of harm; and, especially, the contestability of the category definition and whether it implicates speech, such as political speech, that is ordinarily highly protected. More research is needed for a true assessment of social welfare costs and benefits. This requires greater openness from companies (with a nudge from regulators, if necessary). In the meantime, in these cases, ad hoc, opaque cartelization should not be encouraged.

In any collaboration, adopted tools and standards need to be legitimized through adequate due process, consultation, and reliance on more universally legitimate norms than individual private company judgment. Basing moderation policies on international human rights norms, for example, gives companies a framework for making rights-compliant decisions, along with a globally understood vocabulary for articulating their enforcement decisions to governments and individuals.[128] Homogenization may be less concerning when it is based on independent norms that can be contested and adjudicated outside the four walls of any particular company. This requires adequate oversight.

Legitimizing Cooperation and Avoiding Content Cartels

In a big way, addressing the accountability deficits of cartels requires addressing the accountability deficits of the platforms themselves. This

is an ongoing project beyond my focus here. This section looks at the additional measures that can and should be taken to prevent cartels compounding underlying platform problems. The prescriptions are necessarily general and will need to be adapted to context, but they provide a starting point.

Crucially, the independent oversight that is often part of early plans needs to be meaningfully implemented in the final institutional design. Independent experts and civil society must have access to the information needed to check the processes and outcomes of collaboration. As the GNI example shows, these oversight mechanisms need to be adequately empowered, with the capacity to demand information and raise flags. Acknowledging this necessity, Facebook has recently announced plans to make the GIFCT an independent body[129]—however, without details or a timeline, it is impossible to assess whether this will involve substantive and meaningful oversight. As governments push for cartelization in certain areas, they should build in mandates for independent human rights audits.

Because speech decisions are of such public importance and implicate fundamental rights, legitimation will necessarily include the trite but true call for more transparency. Effective collaboration not only needs to be done but needs to be *seen* to be done. As Sabino Cassese notes about global administrative law, "A fair procedure plays an important role in building social consensus. Process control or voice encourage people's cooperation with authorities and lead to legitimacy."[130] The current procedural and accountability deficits of these arrangements mean that they suffer from legitimacy deficits regardless of whether they are substantively beneficial, because their advantages cannot be verified.

Currently, there is little incentive for companies to be more transparent. Twitter was the only company that released its data on takedowns of Chinese information operations against Hong Kong protestors—as a result, it faced the brunt of scrutiny and criticism about mistakenly identified accounts which other, less open, platforms evaded. But without greater transparency, there can be no accountability. While there are tradeoffs involved in the level of transparency platforms should provide (such as privacy risks or enabling bad actors to game the system), the

status quo is very far from the optimal balance, and the interest that determines what is disclosed is not that of the public interest but that of the companies. What transparency there is often amounts to transparency theater. The GIFCT released its first "transparency report"[131] in the second half of 2019, and at a mere 1,500 words, revealed very little new information.

Transparency mandates should demand metrics that are auditable and also incentivize desirable behavior. For example, when the primary metric reported in the GIFCT report is the number of hashes added to the database, this incentivizes adding hashes but not increasing the accuracy of the hashes added. And because we do not know which platforms are responsible for adding the content or how they use the database to police their individual services, we do not know the power or effectiveness of the individual platforms' participation in the project. Similarly, reports on influence campaign takedowns rarely provide tools to independently evaluate company actions, beyond the salutary news that companies are indeed working together in some capacity and taking some sort of action. But the underlying systems that lead to this collaboration and action, including the level of government participation, remain almost entirely opaque. So too are the biases or blind spots of the technological tools used by cartels—any such technology needs to be subject to independent algorithmic auditing in order to identify and correct any disparate impact.

Accountability also requires access to remediation. If there is centralized censorship, there should be centralized remediation. If a piece of content is taken down because it was in the GIFCT database, for example, then a challenge to this takedown on any individual platform needs to feed back into the GIFCT network so that decisions made collectively do not need to be challenged individually. Similarly, if platforms rely on each other's signals to takedown influence campaigns, there needs to be more information about the extent of this reliance and a process for those who are mistakenly affected to challenge actions taken against them. If done effectively, this could turn an accountability deficit into an advantage, by providing access to remediation for people affected by decisions made on platforms that do not otherwise have the resources or the will to build out an entire appeals process of their own.

Power imbalances in the decision-making structures of these institutions also need to be addressed. Larger platforms may be the source of the technology, but they should not be the source of all decisions. The technical tools still need humans to tell them what to look for, and these decisions should be made through representative and collaborative processes. This will be further facilitated by greater transparency into negotiation processes to ensure that all stakeholders' interests are adequately represented and accounted for. A promising model is that piloted by Cloudflare, which has built in capacity for its clients to calibrate CSAM technology provided by Cloudflare to suit the different needs of different platforms.[132] While such negotiations or adjustments may slow initiatives in the short term or make agreements more difficult or costly to achieve, in the long term they also serve the interests of those companies and governments that hope to relieve public pressure by legitimating these initiatives. For speech decisions, perhaps more than in any other case, openness and accountability are the only way to ensure that decisions made in the name of improving public discourse actually do so.

CONCLUSION

THE MANAGEMENT OF ONLINE PUBLIC DISCOURSE is at a turning point. Platforms are being called on to do more, to do better, and to work together. The choice between ruthless competition and monopoly power is a false one—each poses its own risks to public discourse. Collaboration presents another option, but the task now is to develop institutional designs that legitimize this cooperation and halt the rise of unaccountable content cartels.

NOTES

1 Hany Farid, *Reining in Online Abuses*, 19 Tech. & Innovation 593, 595–96 (2018).

2 *Id*. at 596.

3 Disturbing recent reporting highlighted how much of an issue online CSAM remains, but concerns centered around inadequate resources for law enforcement and NCMEC, companies' failure to comprehensively expand the initiative to video, and the refusal of some services to participate at all, rather than a failure of the underlying theoretical model. *See, e.g.*, Michael H. Keller & Gabriel J.X. Dance, *The Internet Is Overrun With Images of Child Sexual Abuse. What Went Wrong?*, N.Y. Times (Sept. 28, 2019), https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html [https://perma.cc/H3VQ-RF8A]; Michael H. Keller & Gabriel J.X. Dance, *Child Abusers Run Rampant as Tech Companies Look the Other Way*, N.Y. Times (Nov. 9, 2019), https://www.nytimes.com/interactive/2019/11/09/us/internet-child-sex-abuse.html [https://perma.cc/V7KF-K6DL].

4 *See, e.g.*, *The Rise of the European Upload Filter*, Eur. Digital Rights (Jun. 20, 2012), https://edri.org/edrigramnumber10-12the-rise-of-the-european-upload-filter/ [https://perma.cc/429W-RVYY].

5 C.J. Davies, *The Hidden Censors of the Internet*, Wired UK (May 20, 2009), https://www.wired.co.uk/article/the-hidden-censors-of-the-internet [https://perma.cc/R3VU-GWHM]; Emily B. Laidlaw, Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility 116 (2015).

6 *See, e.g.*, Keller & Dance, *The Internet Is Overrun With Images of Child Sexual Abuse. What Went Wrong?*, *supra* note 3 ("The Times found that there was a close relationship between the [NCMEC] and Silicon Valley that raised questions about good governance practices. … This close relationship with tech companies may ultimately be in jeopardy. In 2016, a federal court held that the [NCMEC], though private, qualified legally as a government entity because it performed a number of essential government functions."); Lord MacDonald QC, A Human Rights Audit of the Internet Watch Foundation 23–25 (2013), https://www.iwf.org.uk/sites/default/files/inline-files/Human_Rights_Audit_web.pdf [https://perma.cc/92JJ-V8ZW] (finding that the appeals process and independence of inspections were "insufficiently robust").

7 Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 Notre Dame L. Rev. 1035, 1044 (2018).

8 Ellen Nakashima, *There's a New Tool to Take Down Terrorism Images Online. But Social-Media Companies Are Wary of It.*, Wash. Post (Jun. 21, 2016), https://www.washingtonpost.com/world/national-security/new-tool-to-take-down-terrorism-images-online-spurs-debate-on-what-constitutes-extremist-content/2016/06/20/0ca4f73a-3492-11e6-8758-d58e76e11b12_story.html [https://perma.cc/4EP6-HSF8].

9 *Mass Violence, Extremism and Digital Responsibility: Hearing before the Senate Comm. on Commerce, Science, and Transp.*, 116th Cong. (2019), https://www.commerce.senate.gov/2019/9/mass-violence-extremism-and-digital-responsibility [https://perma.cc/P95M-KXNN]. Nick Pickles of Twitter told the Committee, "We've grown that partnership, so we share URLs. So, if we see a link to a piece of content like a manifesto, we're able to share that across industry. And furthermore, I think an area that after Christchurch we recognized we need to improve, we now have real time communications in a crisis, so industry can talk to each other in real time operationally to say even, you know, not content related, but situational awareness." *Id.*

10 *Id.*

11 *See, e.g.*, Emily Birnbaum, *TikTok Seeks to Join Tech Fight Against Online Terrorism*, The Hill (Nov. 4, 2019), https://thehill.com/policy/technology/468884-tiktok-seeks-to-join-tech-fight-against-online-terrorism [https://perma.cc/EM42-6KEJ].

12 Antigone Davis & Guy Rosen, *Open-Sourcing Photo- and Video-Matching Technology to Make*

the Internet Safer, Facebook Newsroom (Aug. 1, 2019), https://newsroom.fb.com/news/2019/08/open-source-photo-video-matching/ [https://perma.cc/K8BA-S2TC].

13 Citron, *supra* note 7, at 1037–38.

14 *See Global Internet Forum to Counter Terrorism to Hold First Meeting in San Francisco*, Facebook Newsroom (July 31, 2017), https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/ [https://perma.cc/55CJ-X5AZ] ("Building on the work started within the EU Internet Forum and the shared industry hash database, the GIFCT is fostering collaboration with smaller tech companies, civil society groups and academics, and governments.").

15 *Next Steps for the Global Internet Forum to Counter Terrorism*, Facebook Newsroom (Sept. 23, 2019) https://newsroom.fb.com/news/2019/09/next-steps-for-gifct/ [https://perma.cc/2TJN-Y64K] ("GIFCT has made significant achievements since it was founded in 2017, and worked closely with a range of governments, particularly under the auspices of the European Union Internet Forum, but the horrific terrorist attack in Christchurch and the extraordinary virality of the attacker's video online illustrated the need to do even more.").

16 *See* David Uberti, *The German Synagogue Shooter's Twitch Video Didn't Go Viral. Here's Why.*, Vice (Oct. 11, 2019), https://www.vice.com/en_us/article/zmjgzw/the-german-synagogue-shooters-twitch-video-didnt-go-viral-heres-why [https://perma.cc/T3X9-UK95].

17 Eleanor Ainge Roy, *Christchurch Attack: New Zealand Tries New Tactic to Disrupt Online Extremism*, The Guardian (Oct. 14, 2019), https://www.theguardian.com/world/2019/oct/14/christchurch-attack-new-zealand-tries-new-tactic-to-disrupt-online-extremism [https://perma.cc/5DM4-NV84] ("'This will work in a similar way to how we target child sexual exploitation material by working with online content hosts to find and remove harmful content,' [New Zealand Prime Minister Jacinda Ardern] said.").

18 Citron, *supra* note 7, at 1050–51.

19 For an account of how market forces and regulatory and civil society pressure can lead companies to converge on policies and processes of content moderation, *see* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598, 1616–62 (2018).

20 Renee DiResta et al., New Knowledge, The Tactics & Tropes of the Internet Research Agency 42 (2019) (emphasis added).

21 Brian Fishman, *Crossroads: Counter-Terrorism and the Internet*, 2 Tex. Nat'l Sec. Rev. 82, 86–87 (2019).

22 Nika Aleksejeva et al., Atlantic Council, Digital Forensic Research Lab, Operation "Secondary Infektion" 3 (2019) https://www.atlanticcouncil.org/wp-content/uploads/2019/08/Operation-Secondary-Infektion_English.pdf [https://perma.cc/GT7S-ND6Y].

23 *See* Nathaniel Gleicher, *How Does Facebook Investigate Cyber Threats and Information Operations?*, Facebook Newsroom (Nov. 13, 2018), https://about.fb.com/news/2018/11/investigating-threats/#working-with-partners [https://perma.cc/9KHV-4D6U].

24 *See* Victoria Kwan, *Facebook's Ex-Security Chief on Disinformation Campaigns: "The Sexiest Explanation is Usually Not True,"* First Draft (July 9, 2019), https://firstdraftnews.org:443/alex-stamos-interview-disinformation-campaigns/ [https://perma.cc/45QN-FHMX] ("Signs of coordination—of being able to cluster accounts, where you are saying: here are a bunch of actors, perhaps on one platform, or more likely, across multiple platforms, and we have evidence that ties them together—can be a strong source of attribution."). *See also The Lawfare Podcast: Ben Nimmo on the Whack-a-Mole Game of Disinformation*, Lawfare (Nov. 21, 2019), https://www.lawfareblog.com/lawfare-podcast-ben-nimmo-whack-mole-game-disinformation [https://perma.cc/G96K-NN3T].

25 *See, e.g.*, Ali Breland, *Facebook and Twitter Remove Thousands of Accounts Spreading Misinformation from Iran and Russia*, Mother Jones (Jan. 31, 2019), https://www.motherjones.

com/politics/2019/01/facebook-twitter-iran-russia [https://perma.cc/U23Q-NV9M]; Craig Silverman, *An Iranian Disinformation Operation Impersonated Dozens of Media Outlets To Spread Fake Articles*, BuzzFeed News (May 14, 2019), https://www.buzzfeednews.com/article/craigsilverman/iran-disinformation-campaign [https://perma.cc/Z6J7-MRZP].

26 Jessica Brandt & Bradley Hanlon, *Online Information Operations Cross Platforms. Tech Companies' Responses Should Too*., Lawfare (Apr. 26, 2019), https://www.lawfareblog.com/online-information-operations-cross-platforms-tech-companies-responses-should-too [https://perma.cc/E9V5-YQMF].

27 *See, e.g.*, u/worstnerd, Post to r/reddit security, *Suspected Campaign from Russia on Reddit*, Reddit (Dec. 6, 2019), https://www.reddit.com/r/redditsecurity/comments/e74nml/suspected_campaign_from_russia_on_reddit [https://perma.cc/Z7FH-3DZ5].

28 Tony Romm & Ellen Nakashima, *U.S. Officials Huddle with Facebook, Google and Other Tech Giants to Talk About the 2020 Election*, Wash. Post (Sept. 4, 2019), https://www.washingtonpost.com/technology/2019/09/04/us-officials-huddle-with-facebook-google-other-tech-giants-talk-about-election [https://perma.cc/4LC5-975T].

29 2 S. Comm. on Intelligence, 116th Cong., Rep. on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election 80 (2019). [https://perma.cc/QZ6T-25YP].

30 Cat Zakrzewski, *Silicon Valley's Congressman Wants Tech Companies to Collaborate on Disinformation*, Wash. Post: Power Post (June 12, 2019), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/06/12/the-technology-202-silicon-valley-s-congressman-wants-tech-companies-to-collaborate-on-disinformation/5cfffb01a7a0a4030c3fda41/[https://perma.cc/7JUY-47X5].

31 *Id.*

32 Alex Stamos et al., *Combatting State-Sponsored Disinformation Campaigns from State-Aligned Actors*, 43, 48 *in* Securing Am. Elections (Michael McFaul ed., 2019).

33 Nate Persily & Alex Stamos, *Regulating Online Political Advertising by Foreign Governments and Nationals*, 27, 34, *in* Securing Am. Elections, *supra* note 32, at 27, 34. *See also Artificial Intelligence and Counterterrorism: Possibilities and Limitations: Hearing Before the H. Subcomm. on Intelligence and Counterterrorism*, 116th Cong. (2019) [hereinafter *Written Testimony of Alex Stamos*] (written testimony of Alex Stamos, Director, Stanford Internet Observatory), https://homeland.house.gov/imo/media/doc/Testimony-Stamos.pdf [https://perma.cc/8ZPS-FKHU].

34 Paul M. Barrett, N.Y.U. Stern Center for Business and Human Rights, Disinformation and the 2020 Election 20 (2019).

35 *See* Stamos et al., *supra* note 32, at 49.

36 *See, e.g.*, Elise Thomas, *In the Battle Against Deepfakes, AI is Being Pitted Against AI*, Wired UK (Nov. 25, 2019), https://www.wired.co.uk/article/deepfakes-ai [https://perma.cc/75ZT-KB8E] ("Cooperation between the tech giants on the Deepfake Detection Challenge may lay the groundwork for further collective action in the battle against deepfakes. One option might be a system similar to the current Global Internet Forum to Counter Terrorism."); WITNESS & First Draft, Summary of Discussions and Next Step Recommendations from "Mal-uses of AI-Generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening" 26 (2018) (on file with the Knight Institute) ("Potential solution areas discussed included: Better collaboration between platforms on identifying and labelling manipulated content, for example, by running a range of updated screening algorithms on incoming video and audio and sharing this information between platforms.").

37 Zakrzewski, *supra* note 30.

38 Sec'y of State for Dig., Culture, Media & Sport & Sec'y of State for the Home Dep't, Online Harms White Paper 45 (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf [https://perma.cc/3KAY-V3QX].

39 Chandelis Duster, *Kamala Harris Defends Her Push to Get Trump's Twitter Account Suspended*, CNN (Oct. 16, 2019), https://www.cnn.com/2019/10/16/politics/kamala-harris-trump-twitter-suspension-cnntv/index.html [https://perma.cc/4BSS-75CJ] (quoting Sen. Kamala Harris).

40 *Written Testimony of Alex Stamos*, *supra* note 33, at 10 (emphasis added).

41 Perspective, https://www.perspectiveapi.com/#/home [https://perma.cc/3SC2-G7ZB] (last visited Jan. 28, 2019).

42 Jigsaw, *How Latin America's Second Largest Social Platform Moderates More than 150K Comments a Month*, Medium (Aug. 29, 2019), https://medium.com/jigsaw/how-latin-americas-second-largest-social-platform-moderates-more-than-150k-comments-a-month-dfod8a3ac242 [https://perma.cc/72PX-Z946].

43 *See, e.g.*, Katyanna Quach, *Oh Dear ... AI Models Used to Flag Hate Speech Online Are, Er, Racist Against Black People*, Register (Oct. 11, 2019), https://www.theregister.co.uk/2019/10/11/ai_black_people [https://perma.cc/7BCW-6VY3]; Dennys Antonialli, *Drag Queen vs. David Duke: Whose Tweets Are More "Toxic"?*, Wired (Jul. 25, 2019), https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets [https://perma.cc/A8MB-TT94].

44 Courtney Gregoire, *Microsoft Shares New Technique to Address Online Grooming of Children for Sexual Purposes,* Microsoft on the Issues (Jan. 9, 2020), https://blogs.microsoft.com/on-the-issues/2020/01/09/artemis-online-grooming-detection [https://perma.cc/T6NE-4QH4].

45 Nellie Bowles & Michael H. Keller, *Video Games and Online Chats Are "Hunting Grounds" for Sexual Predators*, N.Y. Times (Dec. 7, 2019), https://www.nytimes.com/interactive/2019/12/07/us/video-games-child-sex-abuse.html [https://perma.cc/A2KN-82LD].

46 *Vergecast: Is Facebook Ready for 2020?*, The Verge (Aug. 27, 2019), https://www.theverge.com/2019/8/27/20834965/facebook-alex-stamos-interview-election-2020-vergecast [https://perma.cc/3L9M-NQ8T].

47 W. Va. State Bd. of Educ. v. Barnette, 319 U.S. 624, 642 (1943).

48 I do not take a position on the desirability of breaking up the major tech companies—my focus is on the need to address speech-related harms from social media through other means.

49 Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media (2018).

50 Jillian C. York & Ethan Zuckerman, *Moderating the Public Sphere*, *in* Human rights in the age of platforms 137, 140 (Rikke Frank Jørgensen ed., 2019).

51 Aviv Ovadya, *The "Magical Decentralization Fallacy,"* Medium (Nov. 6, 2018), https://medium.com/@aviv/the-magical-decentralization-fallacy-69b426d16bdc [https://perma.cc/ZNZ2-EZL8].

52 *Id.*

53 Natasha Lomas, *Legislators from Ten Parliaments Put the Squeeze on Facebook*, TechCrunch (Nov. 7, 2019), http://social.techcrunch.com/2019/11/07/legislators-from-ten-parliaments-put-the-squeeze-on-facebook [https://perma.cc/MLP7-5H5B].

54 evelyn douek, *Breaking Up Facebook Won't Fix Its Speech Problems*, Slate (May 10, 2019), https://slate.com/technology/2019/05/chris-hughes-facebook-antitrust-speech.html [https://perma.cc/4LML-HUQB].

55 *See, e.g.*, Gene Kimmelman, *The Right Way to Regulate Digital Platforms*, Shorenstein Center (Sept. 18, 2019), https://shorensteincenter.org/the-right-way-to-regulate-digital-platforms [https://perma.cc/9KR2-CP3Y]; Philip M. Napoli, *What Would Facebook Regulation Look Like? Start With the FCC*, Wired (Oct. 4, 2019), https://www.wired.com/story/what-would-facebook-regulation-look-like-start-with-the-fcc [https://perma.cc/F9CD-DT62].

56 As Chris Hughes worried, in his *New York Times* opinion article calling for the breakup of Facebook, "more competition in social networking might lead to a conservative Facebook and a liberal one." Chris Hughes, *It's Time to Break Up Facebook*, N.Y. Times (May 9, 2019), https://

www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html [https://perma.cc/X5C5-7LBC].

57 This could itself exacerbate other harms. As Jeff Gary and Ashkan Soltani have suggested, "[S]o long as platform profits are reliant on keeping users on-platform as long as possible, controversial and harmful speech will continue to proliferate." Jeff Gary & Ashkan Soltani, *First Things First: Online Advertising Practices and Their Effects on Platform Speech*, Knight First Amend. Inst. (Aug. 21, 2019), https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech [https://perma.cc/3PM2-VC7W].

58 *See, e.g.*, Ellen P. Goodman, *Information Fidelity and Digital Flows*, Knight First Amend. Inst. (forthcoming 2020).

59 Margrethe Vestager, European Commissioner for Competition, *Security and Trust in a Digital World* (Sept. 13, 2019) *in* CCBEInfo (Council of Bars and Law Soc'ys of Eur., Brussels, Belg.), Sept. 2019, at 4, https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/Newsletter/CCBEINFO84/EN_newsletter_84.pdf [https://perma.cc/3Q5M-J4GX].

60 *See* Sec'y of State for Dig., Culture, Media & Sport & Sec'y of State for the Home Dep't, *supra* note 38, at 49 ("Harmful content and behaviour originates from and migrates across a wide range of online platforms or services, and these cannot readily be categorised by reference to a single business model or sector.").

61 April Glaser, *Telegram Was Built for Democracy Activists. White Nationalists Love It.*, Slate (Aug. 8, 2019), https://slate.com/technology/2019/08/telegram-white-nationalists-el-paso-shooting-facebook.html [https://perma.cc/Q8XR-EVEM]; Tess Owen, *How Telegram Became White Nationalists' Go-To Messaging Platform*, Vice (Oct. 7, 2019), https://www.vice.com/en_us/article/59nk3a/how-telegram-became-white-nationalists-go-to-messaging-platform [https://perma.cc/TK93-HNJH] ("The thriving far-right Telegram community is also a reminder that exiling extremists from mainstream social media platforms and forcing their websites, like 8chan, offline, may temporarily inconvenience the movement—but doesn't necessarily fix the problem. And in some cases, it might even make things worse.").

62 Kabir Taneja, *Breaking the Islamic State's Use of Online Spaces Such as Telegram*, Observer Research Found. (Dec. 3, 2019), https://www.orfonline.org/expert-speak/breaking-the-islamic-states-use-of-online-spaces-such-as-telegram [https://perma.cc/4QFK-CQ4B]; Rita Katz, *ISIS Is Now Harder to Track Online—but That's Good News*, Wired (Dec. 16, 2019), https://www.wired.com/story/opinion-isis-is-now-harder-to-track-onlinebut-thats-good-news [https://perma.cc/H9SP-AZTS].

63 *See* Chinmayi Arun, *India May Be Witnessing the Next 'WhatsApp Election'—and the Stakes Couldn't Be Higher*, Wash. Post (Apr. 25, 2019), https://www.washingtonpost.com/opinions/2019/04/25/india-could-see-next-whatsapp-election-stakes-couldnt-be-higher [https://perma.cc/WN2S-YDXY]; Cristina Tardáguila et al., Opinion, *Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It.*, N.Y. Times (Oct. 17, 2018), https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html [https://perma.cc/9YD9-4QS7]; *WhatsApp Both Strengthens and Undermines Nigerian Democracy, Says UK-Nigeria Research Team*, U. Birmingham (Jul. 29, 2019), https://www.birmingham.ac.uk/news/latest/2019/07/whatsapp-both-strengthens-and-undermines-nigerian-democracy-says-uk-nigeria-research-team.aspx [https://perma.cc/L7EV-YSEN].

64 *See* Nilesh Christopher, *TikTok is Fuelling India's Deadly Hate Speech Epidemic*, Wired UK (Aug. 12, 2019), https://www.wired.co.uk/article/tiktok-india-hate-speech-caste [https://perma.cc/9E5Z-EJQC].

65 Alexis C. Madrigal, *India's Lynching Epidemic and the Problem With Blaming Tech*, The Atlantic (Sept. 25, 2018), https://www.theatlantic.com/technology/archive/2018/09/whatsapp/571276/ [https://perma.cc/T2LL-U994].

66 Siva Vaidhyanathan, Antisocial Media: How Facebook Disconnects Us and Undermines Democracy 1 (2018).

67 Julie E. Cohen, *Internet Utopianism and the Practical Inevitability of Law*, 18 Duke L. & Tech. Rev. 85, 89 (2019) [hereinafter *Internet Utopianism*] (emphasis added). *See also* Gary & Soltani, *supra* note 57 ("So long as platform profits are reliant on keeping users on-platform as long as possible, controversial and harmful speech will continue to proliferate.").

68 Lomas, *supra* note 53.

69 April Glaser, *White Supremacists Still Have a Safe Space Online. It's Discord.*, Slate (Oct. 9, 2018), https://slate.com/technology/2018/10/discord-safe-space-white-supremacists.html [https://perma.cc/MXX9-MY64].

70 Google, How Google Fights Piracy 13 (Nov. 2018) https://www.blog.google/documents/25/GO806_Google_FightsPiracy_eReader_final.pdf [https://perma.cc/VC8P-BF4V].

71 Uberti, *supra* note 16.

72 Justin Paine & John Graham-Cumming, *Announcing the CSAM Scanning Tool, Free for All Cloudflare Customers*, The Cloudflare Blog (Dec. 18, 2019), https://blog.cloudflare.com/the-csam-scanning-tool [https://perma.cc/LYU8-K9BF].

73 David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), Rep. on Online Hate Speech ¶50 U.N. Doc A/74/48050 (Oct. 9, 2019).

74 Brandt & Hanlon, *supra* note 26. *See also* Rita Katz, *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.*, Wired (Nov. 20, 2019), https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really [https://perma.cc/8S75-56S8] ("YouTube's victories against ISIS and al-Qaeda must be evaluated in the context of the entire tech industry—including other Google services. … While major platforms like YouTube, Twitter, and Facebook are often widely cited as platforms that have been exploited in ISIS and al-Qaeda's outreach, it is often understated how far these terrorist groups' tentacles reach.").

75 Camille François, Transatlantic Working Group, *Actors, Behaviors, Content: A Disinformation ABC* (2019), https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf [https://perma.cc/2ZTM-MYM7].

76 Alex Stamos (@alexstamos), Twitter (Dec. 7, 2019, 8:10 PM), https://twitter.com/alexstamos/status/1203240448153677826?s=20 [https://perma.cc/7MY9-L3LC].

77 *See, e.g.*, Siddharth Venkataramakrishnan, *Far-Right Extremists Flock to Protest Messaging App Telegram*, Fin. Times (Dec. 16, 2019), https://www.ft.com/content/5e05fc9e-1c35-11ea-97df-cc63de1d73f4 [https://perma.cc/U3DM-RB4B]; Katz, *supra* note 62.

78 Julia Carrie Wong, *Pinterest Makes Aggressive New Move in Fight Against Vaccine Misinformation*, The Guardian (Aug. 28, 2019), https://www.theguardian.com/society/2019/aug/28/pinterest-anti-vaccine-combat-health-misinformation [https://perma.cc/XE2M-32JS].

79 Carmen Fishwick, *How a Polish Student's Website Became an Isis Propaganda Tool*, The Guardian (Aug. 15, 2014), https://www.theguardian.com/world/2014/aug/15/-sp-polish-man-website-isis-propaganda-tool [https://perma.cc/B2GH-5BME].

80 *See Technology Against Terrorism: How to Respond to the Exploitation of the Internet*, Chatham House (July 12, 2017), https://www.chathamhouse.org/event/technology-against-terrorism-how-respond-exploitation-internet [https://perma.cc/TFF2-9DVG].

81 Julia Carrie Wong, *Germany Shooting Suspect Livestreamed Attempted Attack on Synagogue*, The Guardian (Oct. 9, 2019), https://www.theguardian.com/world/2019/oct/09/germany-shooting-synagogue-halle-livestreamed [https://perma.cc/V4XD-6M7W].

82 Sheera Frenkel (@sheeraf), Twitter (Jun. 5, 2019, 9:47pm), https://twitter.com/sheeraf/status/1136449375784251394 [https://perma.cc/56UK-53X7].

83 *See* evelyn douek, *YouTube's Bad Week and the Limitations of Laboratories of Online Governance*, Lawfare (June 11, 2019), https://www.lawfareblog.com/youtubes-bad-week-and-limitations-laboratories-online-governance [https://perma.cc/B3FQ-LM96].

84 *See* Aja Romano, *Apple Banned Alex*

Jones's Infowars. Then the Dominoes Started to Fall. Vox (Aug. 6, 2018), https://www.vox.com/policy-and-politics/2018/8/6/17655516/infowars-ban-apple-youtube-facebook-spotify [https://perma.cc/9E4U-6S77].

85 *See* Kevin Roose, *Facebook Banned Infowars. Now What?*, N.Y. Times (Aug. 10, 2018), https://www.nytimes.com/2018/08/10/technology/facebook-banned-infowars-now-what.html [https://perma.cc/DE62-QLJV].

86 *See* Mary Anne Franks, *The Free Speech Black Hole: Can The Internet Escape the Gravitational Pull of the First Amendment?*, Knight First Amend. Inst. (Aug. 21, 2019), https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment [https://perma.cc/ZT9H-C746].

87 Kate Conger & Nathaniel Popper, *Behind the Scenes, 8chan Scrambles to Get Back Online*, N.Y. Times (Aug. 5, 2019), https://www.nytimes.com/2019/08/05/technology/8chan-website-online.html [https://perma.cc/74YP-UJPE].

88 *See A Conversation with Mark Zuckerberg, Jenny Martinez and Noah Feldman*, Facebook Newsroom (June 27, 2019), https://newsroom.fb.com/news/2019/06/mark-challenge-jenny-martinez-noah-feldman/ [https://perma.cc/3WT6-Q58J]. *See also* Cat Zakrzewski, *Facebook Seeks Outside Help as It Grapples with Content Moderation Problems*, Wash. Post: PowerPost (July 19, 2019), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/07/19/the-technology-202-facebook-seeks-outside-help-as-it-grapples-with-content-moderation-problems/5d30ad0e1ad2e5592fc359a0/ [https://perma.cc/NX5S-JEKG]. For discussion of the oversight board project, *see* evelyn douek, *Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility*, 21 N.C. J. L. & Tech. 1 (2019).

89 Article 19, Social Media Councils Consultation Paper (June 2019), https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf [https://perma.cc/3V3T-LS3B].

90 Emma Llansó, *Platforms Want Centralized Censorship. That Should Scare You*, Wired (Apr. 18, 2019) https://www.wired.com/story/platforms-centralized-censorship/ [https://perma.cc/26QG-QXHY].

91 Exactly how likely is unclear: Platforms try to have it both ways by touting the benefits of collaboration while insisting that inclusion in the GIFCT database does not mean automatic removal by all members, as they all reach "independent" decisions. Without transparency, it is impossible to evaluate these claims. There is no information available about how often members reject other members' determinations or what happens when this occurs.

92 David Kaye, Speech Police 83 (2019).

93 Fishman, *supra* note 21, at 96.

94 *See* Jeff Deutch et al., Electronic Frontier Found., Syrian Archive & Witness, Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content (May 30, 2019), https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content [https://perma.cc/7UZ9-N2PK].

95 Hadi Al Khatib & Dia Kayyali, *YouTube Is Erasing History*, N.Y. Times (Oct. 23, 2019) https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html [https://perma.cc/PAE9-3BJS].

96 Mike Masnick, *EU Tells Internet Archive That Much Of Its Site Is "Terrorist Content,"* Techdirt (Apr. 11, 2019), https://www.techdirt.com/articles/20190410/14580641973/eu-tells-internet-archive-that-much-site-is-terrorist-content.shtml [https://perma.cc/P8RK-LU6P].

97 *See* Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 UC Davis L. Rev. 1149, 1176 (2018).

98 Cat Zakrzewski, *Big Tech Under Pressure to Limit Spread of 8chan and Other Extremist Content*, Wash. Post: PowerPost (Aug. 7, 2019), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/08/07/the-technology-202-big-tech-under-pressure-to-limit-spread-of-8chan-and-other-extremist-con-

tent/5d49f47a602ff17879a188ca/[https://perma.cc/5JQX-NAXF].

99 Farid, *supra* note 1, at 598–99.

100 Robert Gorwa, *The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content*, 8 Internet Pol. Rev. 14 (2019).

101 Laidlaw, *supra* note 5, at 104–08.

102 Letter from Danny O'Brien & Jillian C. York, Electronic Frontier Found., to GNI (Oct. 9, 2013), https://www.eff.org/document/gni-resignation-letter [https://perma.cc/M2NK-ZLDZ] (explaining decision to withdraw from GNI).

103 evelyn douek, *Australia's 'Abhorrent Violent Material' Law: Shouting 'Nerd Harder' and Drowning Out Speech*, Austl. L.J. (forthcoming 2020) (manuscript available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3443220).

104 Liz Woolery, *Three Lessons in Content Moderation from New Zealand and Other High-Profile Tragedies*, Ctr. for Democracy & Tech. (Mar. 27, 2019), https://cdt.org/blog/three-lessons-in-content-moderation-from-new-zealand-and-other-high-profile-tragedies [https://perma.cc/436Z-Z6J7].

105 Uberti, *supra* note 16.

106 Dave Lee, *Baffled Student Tells Twitter: 'I'm Not a Chinese Agent,'* BBC (Aug. 21, 2019), https://www.bbc.com/news/technology-49416617 [https://perma.cc/3D8K-4BYN].

107 *Id.* (quoting Elise Thomas from Australia's International Cyber Policy Centre).

108 Nick Monaco, *Welcome to the Party: A Data Analysis of Chinese Information Operations*, Medium (Sept. 29, 2019), https://medium.com/digintel/welcome-to-the-party-a-data-analysis-of-chinese-information-operations-6d48ee186939 [https://perma.cc/RU67-YUKF].

109 Nathaniel Gleicher, *Removing Coordinated Inauthentic Behavior From China*, Facebook Newsroom (Aug. 19, 2019), https://newsroom.fb.com/news/2019/08/removing-cib-china/ [https://perma.cc/7YTK-9RSD].

110 Shane Huntley, *Maintaining the Integrity of Our Platforms*, Google: The Keyword (Aug. 22, 2019), https://www.blog.google/outreach-initiatives/public-policy/maintaining-integrity-our-platforms/ [https://perma.cc/N2MA-2TB6].

111 Marie C. Baca & Tony Romm, *Twitter and Facebook Take First Actions Against China for Using Fake Accounts to Sow Discord in Hong Kong*, Wash. Post (Aug. 19, 2019), https://beta.washingtonpost.com/technology/2019/08/19/twitter-suspends-accounts-it-accuses-china-coordinating-against-hong-kong-protesters/ [https://perma.cc/5SDA-THHF].

112 Eyal Benvenisti, *Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance*?, 29 Eur. J. of Int'l L. 9, 13 (2018).

113 Julie E. Cohen, Between Truth and Power: The Legal Constructions of Informational Capitalism 220 (2019) [hereinafter Between Truth and Power].

114 *The Call*, Christchurch Call to Eliminate Terrorist & Violent Extremist Content Online, https://www.christchurchcall.com/call.html [https://perma.cc/F44Y-UZ54] (last visited Feb. 6, 2020).

115 Tech Against Terrorism, Case study: Using the GIFCT Hash-sharing Database on Small Tech Platforms, https://www.counterextremism.com/sites/default/files/TAT%20--%20JustPaste.it%20GIFCT%20hash-sharing%20Case%20study.pdf [https://perma.cc/6VA2-BRUC].

116 Fishman, *supra* note 21, at 97.

117 Robert Wright, Why Is Facebook Abetting Trump's Reckless Foreign Policy?, Wired, May 7, 2019.

118 This is, of course, not binary. Cartels can also become sites for powerful governments to seek to have their rules imposed over the wishes of other powerful governments as well as the wishes of other weaker governments. My point is that governments can also leverage cartels to their own ends, and this is rarely to the benefit of smaller states.

119 Between Truth and Power, *supra* note 113, at 218. *See also Internet Utopianism*, *supra* note 67, at 94 ("The powerful global platform businesses that have emerged in the twenty-first

century did not cause any of these changes, but they have proved apt at exploiting them.").

120 Between Truth and Power, *supra* note 113, at 207 ("A giant transnational corporation with operations in many countries can assert interests within all of them and can formulate and advance a unified strategy for furthering those interests.").

121 Letter from Fionnuala Ní Aoláin, Mandate of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, to Mark Zuckerberg, OL OTH 46/2018 (July 24, 2018).

122 Mary Anne Franks, *Censoring Women*, 95 B.U. L. Rev. Ann. 61, 61 (2015).

123 Note, *Cooperation or Resistance?: The Role of Tech Companies in Government Surveillance*, 131 Harv. L. Rev. 1722, 1729 (2018).

124 The staunchest free speech advocates may not accept this proposition. But this is now a minority position: The idea that online speech should be left to the marketplace of ideas alone is a distinctly unfashionable idea, and not one I subscribe to. One can accept that there is both an element of moral panic and risk of overreaction in current debates while also acknowledging that there are areas where effective and responsible content moderation is a necessity.

125 *See, e.g.*, Sarah C. Haan, *Bad Actors: Authenticity, Inauthenticity, Speech and Capitalism*, U. Pa. J. Const. L. (forthcoming) (manuscript available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458795##).

126 Caitlin Petre et al., *"Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility*, 5 Social Media + Soc'y 1 (2019) ("The line between what platforms deem illegitimate algorithmic manipulation and legitimate strategy is nebulous and largely reflective of their material interests.").

127 *Combating Hate and Extremism*, Facebook Newsroom (Sept. 17, 2019), https://about.fb.com/news/2019/09/combating-hate-and-extremism/ [https://perma.cc/9STY-5SPH].

128 *See* David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, *supra* note 73.

129 *Next Steps for the Global Internet Forum to Counter Terrorism*, Facebook Newsroom (Sept. 23, 2019), https://newsroom.fb.com/news/2019/09/next-steps-for-gifct/ [https://perma.cc/RU67-YUKF].

130 Sabino Cassese, *A Global Due Process of Law?*, *in* Values in Global Admin. L. 19–20 (G. Anthony et al. eds 2011).

131 *GIFCT Transparency Report 2019*, Global Internet F. to Counter Terrorism, https://gifct.org/transparency/ [https://perma.cc/CY8K-DLXC] (last visited Feb. 4, 2020).

132 Paine & Graham-Cumming, *supra* note 72.

About the Author

EVELYN DOUEK is an S.J.D. candidate at Harvard Law School and affiliate at the Berkman Klein Center for Internet & Society. Her scholarship focuses on international and transnational regulation of online speech and content moderation institutional design. Her research has appeared in numerous outlets, including the *Atlantic*, *North Carolina Journal of Law and Technology*, and *Federal Law Review*. She also blogs at Lawfare. Before joining Harvard to complete a Master of Laws, Douek clerked for Chief Justice Susan Kiefel of the High Court of Australia and worked as a corporate litigator.

Acknowledgments

Many thanks to Alex Abdo, Rafe Andrews, Katy Glenn Bass, Brenda Dvoskin, Jack Goldsmith, Robert Gorwa, Jameel Jaffer, Thomas Kadri, Kate Klonick, Emma Llansó, Martha Minow, Amre Metwally, the participants at a Berkman Klein Center roundtable discussion for helpful comments and conversations in the development of this piece, and Sarah Guinee, Tiffani Burgess, Jun Nam, and the staff at the Knight First Amendment Institute for their careful editing. Their contributions show the upsides of healthy collaboration, but all errors remain my own.

© 2020, Evelyn Douek.

About the Knight First Amendment Institute

The Knight First Amendment Institute defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. Its aim is to promote a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org
Design: Point Five
Illustration: © Edmon de Haro



KNIGHT
FIRST AMENDMENT
INSTITUTE

at Columbia University