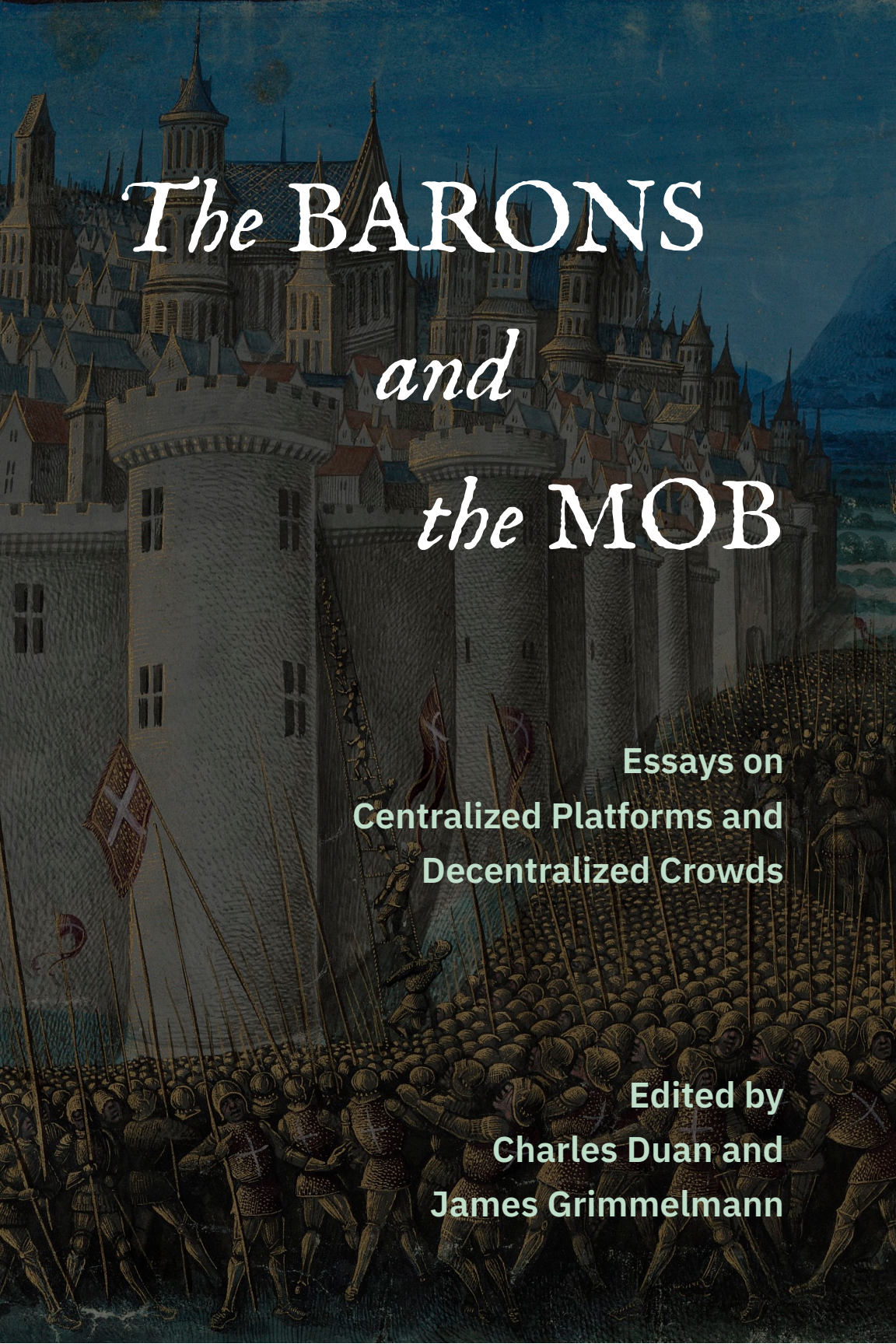


The BARONS *and the* MOB

Essays on Centralized Platforms and Decentralized Crowds

Edited by
Charles Duan and
James Grimmelmann

The Barons *and the* Mob

The BARONS *and the* MOB

Essays on
Centralized Platforms and
Decentralized Crowds

Edited by
Charles Duan and
James Grimmelmann

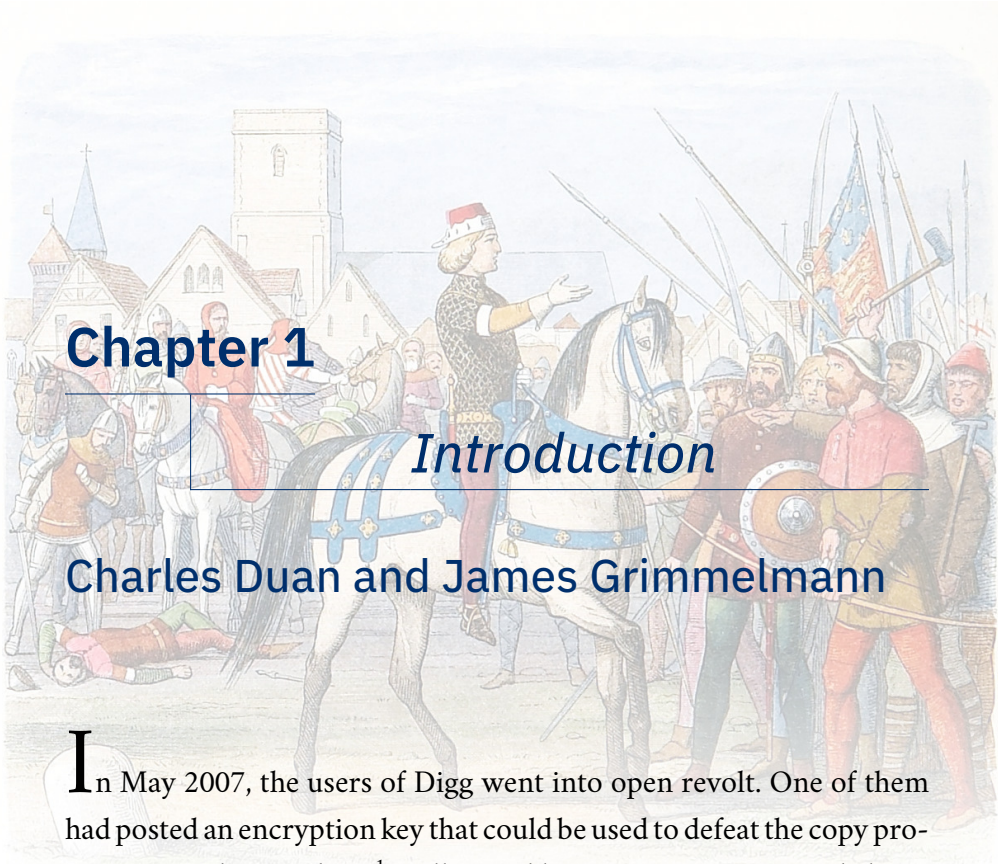
Cornell Tech Research Lab in Applied Law + Technology
2 West Loop Road, New York, NY
2024

This book is typeset in Cochineal, with headings in IBM Plex Sans.

Cover graphic: Jean Colombe, *Le siège d'Antioche* (1097–1098), in *Passages d'outremer* by Sébastien Mamerot (c. 1474).

Chapter graphics: Engravings by Edmund Evans, in *A Chronicle of England: B.C. 55–A.D. 1485* by James William Edmund Doyle (1864).

Contents



Chapter 1

Introduction

Charles Duan and James Grimmelmann

In May 2007, the users of Digg went into open revolt. One of them had posted an encryption key that could be used to defeat the copy protection on Blu-Ray discs.[1] Hollywood lawyers sent a cease-and-desist letter, and the administrators of the popular link-sharing platform removed the post. Annoyed users kept on reposting it, until Digg banned their accounts to make it stop—but it didn't stop. Digg users flooded the site with the key, until, at one point, every post on the site's front page was about it. Eventually, Digg's leaders threw in the towel. The founder wrote a blog post whose title was the key had tried so hard to suppress, and explained:

> But now, after seeing hundreds of stories and reading thousands of comments, you've made it clear. You'd rather see Digg go down fighting than bow down to a bigger company. We hear you, and effective immediately we won't delete stories or comments containing the code and will deal with whatever the consequences might be.
>
> If we lose, then what the hell, at least we died trying.[2]

In 2023, history rhymed with itself on Reddit, the popular news-and-discussion site. In preparation for a rumored IPO, the site started charging developers to access its previously-free API. The prices were so high it was impossible for popular third-party apps to stay in business. The volunteer user-moderators who ran many of Reddit's most popular communities (called "subreddits") were outraged, because many of them depended on third-party apps to keep up with the burden of content moderation. So they too went into revolt. Thousands of subreddits simultaneously went dark, voluntarily shutting off access to their contents to protest the API fees.[3]

Like Digg's management, Reddit's management initially cracked down: threatening moderators with suspension from their moderation roles unless they turned their subreddits back on again.[4] Many of them acceded to the company's demands. But some stuck to their demands, finding other ways to protest. The r/aww subreddit—described as "the online hub of cute animal pictures"[5]—took a different approach. By community vote, r/aww adopted a rule requiring that all content feature John Oliver, the British host of the comedy news show *Last Week Tonight*.[6] Other subreddits went further in their creative disobedience: Several designated themselves as "not safe for work" (despite still prohibiting pornography and other offensive content), precluding Reddit from displaying advertisements on and thereby monetizing those subreddits.[7] But in the end, the uprising failed. Reddit replaced the moderators of subreddits who refused to toe the line. Users, on the whole, mostly kept using the site. The IPO went ahead in March 2024, with Reddit's owners firmly in charge of the site's operations. Reddit didn't bow down; it was the rebellious moderators who went down fighting.

The Digg disruption and the Reddit rebellion demonstrate the conflict between the two great sources of power on the Internet: the cen-

tralized **platforms** that control the infrastructure of online communities, and the decentralized **crowds** of users who come together in them. Both of these groups think that these virtual spaces are theirs by right of ownership, and both of them wield distinctive forms of power.

Platforms have direct control over the technical infrastructure, huge quantities of data, the ability to act quickly and decisively, and often access to large sums of money. They can make technical changes that radically transform the architecture of an online space, and make them stick. They can remove content, ban users, rewrite history, and remake a platform by rewriting its code.

But the users who create, communicate, and do commerce in these spaces have their own forms of power. For one thing, they can leave—at the end of the day, a platform without users is a ghost town. But they can also stay and organize, as on Digg and Reddit—work together to state their case, creatively and pervasively. Users are are neither an undifferentiated mass nor isolated automata. Instead they come together in crowds, groups bound together by collective identity, norms, and purpose.

When platforms meet crowds, these two great forces of the Internet shape and reshape each other. Platforms code features to orient and constrain user behavior, but the street finds its own uses for things, and crowds of users teach each other to evade and repurpose those features. Popular user innovations—like the original Twitter's

When platforms meet crowds, these two great forces of the Internet shape and reshape each other.

hashtags—become official platform features, and platforms seek ways to give their users a sense of proud collective identity. (Reddit, for example, gave its heaviest users the ability to buy shares in its IPO.) The history of the modern Internet is the history of the relationship of platforms and crowds.

Despite its importance, little attention has been paid to the intricate balancing act between platforms and crowds. Legal proposals to deal with platform power—antitrust, privacy, freedom of speech—often proceed along an entirely separate track than legal proposals to deal with online crowds—harassment, meme stocks, and the viral spread of misinformation. And yet the two are profoundly related. You cannot understand platforms without understanding the crowds who congregate on them. You cannot understand crowds without understanding how platforms shape their behavior. Proposals to regulate one or the other in isolation miss more than half the story.

This report is an initial contribution towards understanding platforms and crowds together. It presents the outputs of a virtual workshop, conducted on July 28 and August 4, 2022, bringing together scholars and industry experts from social science, law, and technology. The workshop aimed to address a set of related questions:

- What is the nature of online crowds? How do they form, grow, behave, and interact, both on a platform and across platforms?

- How do crowds wield influence, not just online but also over politics, financial markets, and consumer behavior?

- Are there "good" and "bad" online crowds? What distinguishes desirable crowd activism from mob harassment, and what role do notions of "authenticity" play in that distinction?

- How do crowds and crowd behavior intersect with the affordances and features of online platforms? How do content moderation policies and tools, typically designed around individual behavior, operate with respect to coordinated crowd activity?

- Most importantly, how will proposed legal reforms affecting platforms interact with crowds? What unexpected or unintended consequences might arise if such reforms do not account for the nature of crowds and the crowd–platform dynamic?

At the close of the workshop, we invited our academic participants to reflect on the conversations and to identify important themes about platforms and crowds. This report gathers their views. Together, the pieces offer a compelling introduction to the complexities of online crowds and the importance of understanding their nature as part of meaningfully effective efforts toward online platform regulation.

First, **Charles Duan and James Grimmelmann** taxonomize the relationships between platforms and crowds. They sketch the different ways that platforms and crowds have power over each other; the ways that they manage speech, data and money; and the ways in which legal interventions can reshape platform-crowd relationships. Their emphasis is on the interconnectedness of all things; pulling on one thread can unravel some surprising stitches.

To explain how online crowds form, **Jessica L. Beyer** identifies two key ingredients: shared interests and a common space. Shared interests can arise from a variety of sources, as Beyer explains: Existing identity ecosystems such as political affiliations, common information deficits such as new parents seeking advice, or offline leaders bringing adherents online. The nature of the online space can then influence how a crowd of similarly-interested users produces collective goods, engages in political activism, or otherwise takes action en masse.

Building upon this theory, **Rebecca Tushnet** explores media fandom as an especially potent form of online crowd. Organized groups of fans can generate remarkable works of creativity and also be powerful sources of political action, and Tushnet documents both positive and negative coordinated activity of fandom crowds. In particular, and based on her own experiences operating an online fanfiction service, she describes how fandom crowds take advantage of the technical affordances of platforms—weaponizing content moderation tools and hopping across different services, for example.

Even crowds engaged in the most unethical behaviors nevertheless have ethics of their own—that is the perhaps unexpected thesis of **Gabriella Coleman**'s chapter. Coleman traces the norms of anonymity and anti-celebrity on the so-called "offensive internet." Originally the result of technical limitations of platforms like 4chan, anonymity bound members of the mid-2000s crowd of users to ethics of communal authorship and shunning of individuality, which informed the Anonymous hacktivism movement. Coleman also explains more recent far-right activism as an evolution, and to an extent a fragmentation, of these anonymity norms.

All markets are crowds, **Finn Brunton** observes, and understanding their dynamics as crowds can explain many of the perhaps odd dynamics of online (and offline) markets. Like other crowds, markets can be made to move by calculated efforts of influencers. But members of a market-crowd know this, so every market participant is simultaneously a member of the crowd, trying to predict how the crowd will be influenced, and possibly trying to influence the crowd. Brunton explores the consequences of this self-referential dynamic, especially in view of "augmented manipulation" through bots.

Evelyn Douek comments on a growing trend in content moderation. Rather than merely evaluating individual content items,

platforms are focusing on the structure and actions of crowds as the basis for moderation decisions. They take this moderation-by-association approach both to be efficient and to respond to sophisticated crowds that flood platforms with messages that are individually only mildly objectionable, but in aggregate capable of sowing disinformation or causing other harms. Given the countervailing value of freedom of association, though, Douek calls for greater transparency into how platforms are moderating content based on group association.

Alice Marwick dives further into one component of associational moderation: platform rules about "inauthentic" behavior. Such rules, adopted on many major platforms today, in theory target the disinformation campaigns and other harmful crowd behavior. However, Marwick explains that "authenticity" is a slippery term, which arguably condones paid influencer endorsements but disallows identity-affirming names. Marwick's essay points to the difficulty of formulating rules for managing online crowds, especially in view of Tushnet's observations on how crowds can manipulate rules to their benefit.

With respect to the problem of misinformation, **Srijan Kumar and Bing He** explore leveraging online crowds for fact-checking. Through their experimentation and research, Kumar and He find that crowds on social media can quickly respond to and counter misinformation. Crowd-powered misinformation correction faces critical limits, however: Correction efforts he correction messages are often impolite and themselves lacking evidence, the targets of correction are chosen inconsistently, and coordinated crowds can manipulate the correction tools to boost misinformation, they find. Accordingly, they propose developing assistive tools to empower fact-checking crowds.

Nikolas Guggenberger reviews the policy landscape of online platform regulation reforms. Guggenberger explains that lawmakers

are concerned for the future of public discourse, individual interests, and technological innovation due to the dominance of key social media firms. He reviews the categories of reforms on the table, and considers how they may interact with platforms' treatment of crowd behavior. Interoperability between platforms, for example, could alleviate the rise of extremism on otherwise isolated services; it may also prompt a need to rethink the structure of content moderation.

In evaluating these platform regulation options, **Paul Ohm** invites us to consider crowd dynamics as a window into designing more socially beneficial online services. Ohm points to podcasting and Reddit as models of online businesses that build identifiable communities and then make money through "contextual advertising" to communities as a whole. He argues that this business model is healthier than traditional social media that depends on privacy-invasive individual surveillance, and suggests ways that traditional social media can shift toward contextual advertising.

From these pieces, a picture emerges of key lessons about online crowds, lessons that should shape the ongoing conversation about platform power and regulation. The broadest of these is simply that crowds operate across platforms and often use different platforms to different ends—Facebook to recruit, 4chan to opine, private messaging to organize. Reforms that standardize features, through interoperability for example, may reduce the attractiveness of this cross-platform activity but may also enhance crowds' ability to reach and influence new audiences. Limiting platform size may reduce a crowd's sphere of influence on that platform but may encourage the crowd to further engage across platforms.

Online crowds range from those with well-defined, hierarchical leadership to decentralized ones where everyone is simultaneously influencer and influenced. Platform regulations will have to deal with

Platform features can have key effects on the types of crowds that form.

this full spectrum of crowd leadership arrangements. Vertical separation between content hosting and production, for example, may create vacuums of power that could create new opportunities for crowd leaders to rise. It could also give decentralized crowds an advantage over centralized structures that leaned on integrated platforms. Additionally, data privacy and advertising regulations might make it more difficult for platforms or regulators to discern crowd dynamics.

Platform affordances shape the types of online crowds that form and the norms that crowds embrace, but they also are tools that crowds can exploit, sometimes to undesirable ends. Reforms that alter the features or operations of platforms need to be considered through both of these lenses. A must-carry requirement for platforms could foster a greater diversity of crowds or temper misinformation within a crowd, but crowds could also leverage the must-carry requirement as part of automated influence campaigns. The same double-edgedness could arise from mandatory removal regimes.

While the legislative conversation has largely focused on substantive regulation of content, our workshop conversation showed that seemingly non-substantive architectural platform features can have key effects on the types of crowds that form. A lack of real-names policies can create ethics of anonymity and collectivity; the unauthenticated pull architecture of RSS determines how podcasters advertise to communities. In view of this, ideas such as introducing friction into online posting perhaps merit greater study, since they may have

nonobvious effects on the types of crowds that form and substantive content that finds its way online.

Crowds often exist to produce information goods, either for their own benefit or to attract outsiders to further the crowd's aims. That information can be immensely valuable: Collective wisdom can be a powerful tool for responding to misinformation, and fan works contribute to creative literature. But crowd-generated information can also be destructively manipulative or harmful. Regulations must deal with the difficulty of distinguishing these positive and negative forms of crowd-produced information. In particular, measures of "authenticity" directed to deanonymization of online speakers appear at best to be an incomplete proxy for information quality, in view of the diversity of information goods that crowds produce.

Our workshop was titled "The Barons and the Mob" to capture the dual centers of online power that we hoped to explore. Perhaps less than coincidentally, the subsequent Reddit controversy picked up on the same motifs of feudalism, with the company's CEO Steve Huffman calling the protesting subreddit moderators the "landed gentry"[8] and the moderators calling Huffman a "wealthy baron" in response. There are certainly similarities between the phenomena that this report characterizes and the often tumultuous relationships between lords and commoners throughout history. But the overarching finding of this project has been that the online environment makes the relationship between platforms and crowds more complex. The mobs of medieval England could not incite millions of followers with a couple of viral posts; the barons could not rewrite the fortunes of millions of users by changing a line of code. We hope that the social, economic, and political effects between online crowds and platforms, illuminated by the thoughtful essays in this report, can deepen the conversation about shaping our future technological environment.

We hope that the ideas in these pages will be fruitful to all who ponder the Internet, and what will become of it.

Charles Duan and James Grimmelmann
July 2024

Notes

1 *See* Brad Stone, *In Web Uproar, Antipiracy Code Spreads Wildly*, N.Y. Times, May 3, 2007, at A1, https://www.nytimes.com/2007/05/03/technology/03code.html.

2 Kevin Rose, *Digg This: 09-f9-11-02-9d-74-e3-5b-d8-41-56-c5-63-56-88-c0*, Digg Blog (May 1, 2007), https://web.archive.org/web/20070504054516/http://blog.digg.com/?p=74.

3 *See* Michael Levenson, *Reddit Communities Go Dark to Protest New App Policy*, N.Y. Times, June 13, 2023, at B3, https://www.nytimes.com/2023/06/12/business/media/reddit-subreddit-blackout-protest.html.

4 *See* Jay Peters, *How Reddit Crushed the Biggest Protest in Its History*, The Verge (June 30, 2023), https://www.theverge.com/23779477/reddit-protest-blackouts-crushed.

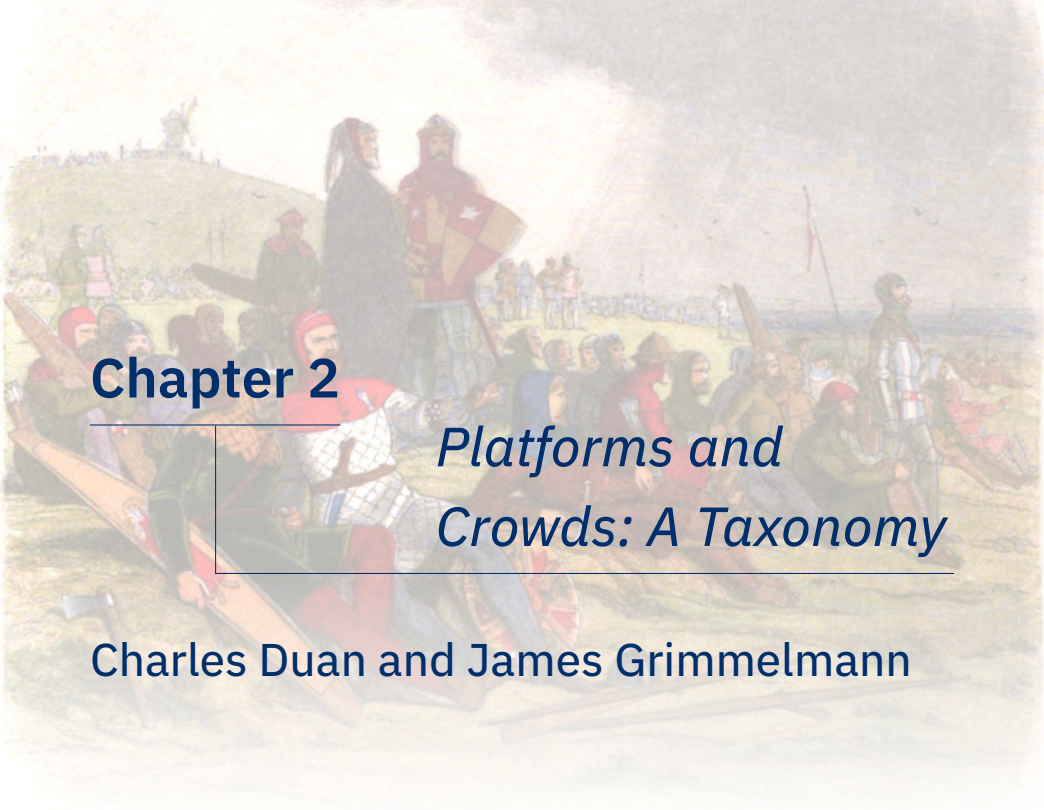
5 *See* James Meese, *"It Belongs to the Internet": Animal Images, Attribution Norms and the Politics of Amateur Media Production*, *in* 17 M/C J. No. 2 (Feb. 24, 2014), https://journal.media-culture.org.au/index.php/mcjournal/article/view/782 (capitalization adjusted).

6 *See* AutoModerator, *Henceforth, /r/aww Will Only Feature John Oliver, Chiijohn, and Their Lookalikes Being Adorable!*, r/aww (Reddit June 18, 2023), https://www.reddit.com/r/aww/comments/14cgp6d/henceforth_raww_will_only_feature_john_oliver/; Tim Marcin, *Popular Subreddits End Their Reddit Protest with *Only* Pictures of John Oliver*, Mashable (June 18, 2023), https://mashable.com/article/subreddits-john-oliver-reddit-protest-return.

7 *See* pics-moderator, *On The State of /r/PICS: Profanity, Offensive Content, and An Open Letter*, r/pics (Reddit June 26, 2023),

https://www.reddit.com/r/pics/comments/14jl5n8/on_the_state_of_rpics_profanity_offensive_content/; Jay Peters, *Reddit Demands Moderators Remove NSFW Labels, or Else*, The Verge (July 7, 2023), https://www.theverge.com/2023/7/6/23786474/reddit-nsfw-moderator-protest-final-warning.

8 *See* David Ingram, *Reddit CEO Slams Protesters, Says He'll Change Site Rules*, NBC News (June 15, 2023), https://www.nbcnews.com/tech/tech-news/reddit-protest-blackout-ceo-steve-huffman-moderators-rcna89544.



Chapter 2

Platforms and Crowds: A Taxonomy

Charles Duan and James Grimmelmann

On the Internet, platforms and crowds are defined by each other. What makes a platform *a platform*, and not just infrastructure, is that it it brings users together and facilitates their interactions. What makes a crowd *a crowd*, and not just a collection of unrelated individuals, is their collective behavior as an entity. Offline, crowds can assemble anywhere there is space for them, but on the Internet they typically come together on a platform—and a platform only becomes one when they do. Platforms make crowds, and crowds make platforms.

Platforms, Crowds, and Power

Platforms have power over crowds. Most directly, they have technical power. They can use that power to constrain: to constrain who can be on the platform at all, what content they can post, and who they can share it with.[1] But they can also use that power to enable (which is the flip side of constraining). Platforms allow users to connect, share, and

assemble. The specific affordances of different platforms enable different types of crowds.[2] Reddit's system of subreddits encourages long-term communities of shared interests, from carpentry to cosplay; GoFundMe's campaigns encourage short-term outbursts of focused charity; TikTok's algorithmic curation seems almost designed to inhibit the formation of coherent crowds with stable identities.

As these examples suggest, platforms also have a more limited kind of social power over the crowds that use them. They can try to shape the norms both of their users in general and of specific groups of users. Through content moderation, user-interface design, and public messaging, platforms try to nudge their crowds towards behaviors the platforms prefer.[3] Do spend money to tip livestreamers; don't harass other communities. Do engage in elaborate rituals of creative collaboration; don't share pornography. Do talk about celebrity drama; don't talk about politics. They may not always succeed, but platforms try to craft their crowds.

It may seem that the power is entirely one-sided—platforms can always pull the plug—but that misses the mutual dependence that characterizes their symbiotic relationship. Crowds always ultimately can exit a platform, and if they do, the platform can lose both its reason for being and the financial support it requires to keep on being.[4] This is not always a threat that can be made lightly—they may or may not have anywhere else to go, and it may or may not be easy to get ev-

> Crowds typically come together on a platform—and a platform only becomes one when they do.

eryone to move together—but it does ultimately give crowds of users substantial leverage over a platform. This was the threat underneath the shenanigans at Digg and Reddit: undo this policy change or we'll quit *en masse*. The threat worked at Digg, and failed at Reddit. But it was made at both, and both platforms understood it as such.

Crowds also have voice at platforms. Their posts and other activity are the most visible and literal form of voice, even when they are not explicitly directed to the platform. Moderators speak of online community "health," and platforms are often attentive to what their users are satisfied and dissatisfied with. The reflexive quality of intra-crowd discourse is often concerned with—often obsessed with—the crowd's norms and its relationship to the platform.[5] They can speak out and act in other ways, too, some of them disruptive, from protests to boycotts to denial-of-service attacks.[6] All of these have a core *collective* dynamic that is characteristics of crowd activity.

In other settings, coordinated crowd behavior is completely essential to the functioning of a platform. A blockchain is literally held together by its participants' consensus; they collectively agree on the state of the blockchain, and act in coherent ways consistently with an accepted protocol. They do so from a mixture of individual incentives (algorithmically mediated, within protocols that are game-theoretically designed to encourage them to cooperate) and social solidarity (including shared goals and an understanding that those goals' success depend on collective cooperation). Their platform—the blockchain that brings them together—is what the crowd decides it is, and blockchain protocols are crafted so that crowd dynamics prevail over individual objections.

Wikis, too, organize crowds for core governance and moderation work. Here, participants' motivations depend on understanding themselves as part of a collective with a shared purpose that is likely to suc-

ceed, and their norms are directed to preserving the conditions under which the collective continues to share that purpose.[7] The collective authorship in a wiki, or the compilation of texts in a fansite, reflect the contributions of many, partially submerged in the collective and partially not. Managers of the platform, those who pay for and run the computer hardware, may have little or nothing to do with the content of the wiki. The collective organization of these sites—ranging from absurdly logical to utterly chaotic—too is a reflection of crowds using platform affordances to build structures of meaning.

And at times, crowds transcend any single platform. Fans of a YouTube star may donate on Patreon, discuss on Reddit, edit Wikipedia, and plan live gatherings on Facebook. These multivalent crowds further shift power away from platforms and toward themselves. They can pick and choose the features and affordances of each platform to their liking—a closed, anonymous one for planning coordinated efforts, and an open, public one for recruiting new members, perhaps.[8] Exiting an unfavorable platform becomes easy, and the platforms must cater to crowds in order to keep them. Competition among platforms can be a good thing, but it may also enhance the excesses of mob mentality or leave less-connected users in the dust. Because crowds are not intrinsically tied to one platform, they wield power to shape all platforms.

Platforms, Crowds, and Speech

Crowds play a central role in the spread of information and ideas through platforms. Sharing is a social activity, and cannot be understood apart from the audiences and communities with whom something is shared. Some of these audiences are preexisting ones, defined by explicit sharing mechanisms at the platform level. Others are meant

for publics at large, communities created or imagined by users who share. And in between are implicit communities defined by nebulous boundaries of familiarity, privacy, and trust.

Crowds come together to push memes and messages, from coordinated brigading attacks to lighthearted jests.[9] No one who recorded or listened to a Wellerman cover in 2022 or 2023 knows who all of the other participants were, but there was a musical crowd all the same. The surging froth of a QAnon message board is a collective endeavor, with users joined in a chaotic collective decipherment of a secret reality. Every platform has its own crowd-level controversies; political anger at TikTok has been fueled by viral stunts and political memes that circulate there. These phenomena are treated as worthy of recognition and response because they rise above the level of mere content—they reveal the stirrings of larger entities with complicated motivations and striking capabilities for action. Here there be crowds.

One important dynamic, which is both typical of some platforms and also widely crosses platform boundaries, is coordinated harassment. *See* Danielle Keats Citron, *Hate Crimes in Cyberspace* (2014), https://www.hup.harvard.edu/catalog.php?isbn=9780674659902. A crowd settles on a target—sometimes diffusely but often as a result of focused targeting by influential users—and then members direct anger and abuse at them. The crowd acts as a magnifier; a torrent of abuse from numerous members can be far harder to bear than a trickle from one user. The crowd also helps to shield its members from accountability, through safety in numbers and obscuring individuals' roles and identity. The crowd, crucially, also helps to shape its own role—participants encourage each other on and help themselves feel a sense of righteous belonging.[10]

At the same time, the definition of what even is crowd activity is highly contested. The category of "coordinated inauthentic behav-

ior" is defined in terms of coordination—something characteristically done by a crowd—but also in terms of authenticity.[11] It is crowd-like, but in some sense does not count as though a real crowd were doing it. Many influence campaigns are waged by a small number of people trying to pass themselves off as large crowds, through the use of sock puppets, bots, and AI.[12] A crowd has a perceived power and legitimacy that an individual may lack.

Platforms, Crowds, and Data

Platforms and crowds regularly struggle for control of data. Users generate data, simply by acting and interacting online. Platforms collect and aggregate that data, for purposes including targeted advertising and to predict user behavior. For platforms that make their services available to users for free, data analytics and advertising are core to making those platforms work.

Crowds intersect with data and data-based platforms in several ways, some less straightforward than others. Most obviously, crowds themselves are data: users become a subject of quantifiable prediction when they act in consistent, correlated ways. The distinguishability of a coherent group of users—members of a crowd—sets them apart from others users—non-members—in ways that the platform can identify and act on.[13] Surveilling users becomes a way to assign them to groups for purposes of monetization and control. Targeted advertising panels—groups of users who share features of interest to advertisers—are only the most obvious example.

The Cambridge Analytica scandal is an interesting case study here. It involved two steps. First, researchers exfiltrated data on millions of users by having them take a personality survey and grant the app permission to view their contacts. That is, they took advan-

tage of the networked links among users to obtain information on their social connections, viewing users as embedded in groups. Then, Cambridge Analytica itself targeted political advertising to them as members of a different set of groups—political microdemographics who could potentially be induced into changing their voting behavior.

In the limit, however, targeted advertising and analytics cause the crowd to drop away. Compared with clustering and classification methods that assign users to demographic and behavioral groups, approaches that rely on individual targeting ultimately speak to individuals as individuals, with messages that are uniquely appealing—or uniquely manipulative.

By contrast, crowds can also defy traditional platform data models. The power of crowds to magnify viewpoints and emotions can lead members of a crowd to behave in ways entirely divorced from their non-crowd lives. A person who is quiet and unassuming in ordinary life may have an online persona as a brilliant fanfiction writer, a powerful meme stock influencer, or a virulent troll.[14] That crowd mentality can change people's data-generating behaviors offers yet another motivation for the increasingly important study of contextual privacy. Indeed, crowds often embrace individual anonymity as part of melding the members into a larger collective.

And there may be a more insidious interaction. When a user of a platform acts as a member of a crowd, it is that user's intensified, crowd-driven persona that generates data. Platforms, relying on that data to present advertisements and content, may push the user further into the world of the crowd. On top of that, crowds may intentionally manipulate data, tricking platforms into thinking that movements are bigger or more widely accepted than they actually are. Contemporary debates over election manipulation, disinformation, filter bubbles, and

algorithmic curation can thus only be complete by accounting for the nature of crowds.

Finally, crowds may offer an escape hatch to a data-driven platform industry. Crowds are monetizable, as any Instagram influencer knows. And often, they are monetizable without the invasive surveillance and data gathering that platforms typically use today.[15] That Instagram influencer can make a quirky jacket sell out simply by sending followers to the store, without knowing every detail about the lives of those followers. The ability of crowds to engage in commerce and move markets represents an alternative, if not an antidote, to platform monetization through data.

Platforms, Crowds, and Money

Of course, that commercial power of crowds is not without its own consequences. When a crowd engages in economic activity, pinning responsibility on a single actor is difficult. A crowd of enraged investors sent GameStop's stock unexpectedly soaring, prompting a halt on trading.[16] But who was to blame: Reddit, the platform where the stock went viral, or Robinhood, the platform of choice for the traders? Both played a part in the crowd's impact, but neither was an essential point of control.

The coherence of crowds also has distinctive consequences for platform economics. First, it is a source of positive network effects: users receive value (social, financial, or both) from each others' participation. We are accustomed to talking about this as a feature of the platform itself, but it is really about common membership in an interacting community of users, which the platform just happens to facilitate.[17] This value is initially enjoyed by users, and of course platforms try to capture some of it for themselves. The startup conventional wisdom—

first pursue growth, and then try to monetize it—reflects this idea, that the first and most important characteristic of a crowd is its size.

Once a crowd has come together on a platform, these same network effects tend to keep them there. Any individual user who departs for another platform gives up the community on the first one; even if the new platform is superior in other ways, there is no one to share it with. This is the familiar phenomenon of lock-in; social media in particular are extraordinarily stable because collective user dynamics give them immense inertia. Even when there are no other technical or financial barriers to entry, the first platform's popularity is self-perpetuating. Indeed, a popular platform will pull users from a less-popular one; the ability to offer the presence of others is a significant competitive advantage. Crowds help to create winner-take-all dynamics in platform competition.

This is one reason that interoperability looms so large in discussions of platform economics.[18] The ability to move between platforms while remaining part of the same communities reduces the competitive advantage of popularity as such. Interoperability allows crowds to migrate from one platform to another gradually and over time, avoiding the collective-action problem in which the earliest adopters of competing platform are punished for it via isolation, even if most users would be better off making a switch. Interoperability also allows crowds to persist over multiple platforms, retaining their collective identity in a way that is not tied down to a single platform. The network as a whole becomes their home, rather than the infrastructure of any one provider.

At the same time, crowd dynamics also complicate the story of interoperability and migration. A crowd can act collectively and discontinuously, taking coordinated action in response to situations that implicate their shared values. One way of looking at the mass migrations

off of Twitter following its sale to Elon Musk in 2022 is that numerous large crowds of users collectively decided that remaining on Twitter was an intolerable option.[19] Crowd-level social animus towards Musk played a significant role in their decisions; atomized individual users might not have made the leap to the same extent. (And conversely, its new owner benefitted from the influx of pro-Musk crowds, including Tesla fans and diversity skeptics.) These crowds settled on alternative homes including Mastodon, Bluesky, Threads, and Post and it is striking how different groups of users scattered to different platforms. Crowds push and jostle among platforms, rather than flowing smoothly.

Platforms, Crowds, and Law

The diversity of ways that platforms and crowds interact have significant implications for laws that try to shape their behavior. It is not just that many bodies of law affect platforms and users (although they do). And it is not just that these bodies of law may affect each other when they all try to regulate the same behavior (although they do). The essential challenge is that laws regulating platforms will also affect crowds, and vice-versa. Sometimes, the most effective way to change the behavior of a platform or a crowd is by way of the other. And thinking through the likely consequences of an intervention requires thinking about the resulting crowd dynamics. Consider just a few examples:

- Abusive communications are just the tip of a spear that has been set in motion by many other factors. Focusing on the legality of those communications can obscure the way that toxic dynamics of networked abuse depend heavily on platform architecture, including sharing affordances, content moderation policies, and user privacy.[20] All of these are potentially subject to

platform control—and to regulation—but complex crowd psychology means that the effects of interventions can be surprising and nonlinear. At the same time, the diffuse and multi-step relationship between platform affordances and resulting abusive behavior means that it can be difficult to draw clearly causal connections and to specify standards of design and behavior for platforms.[21]

- Privacy laws take account of the ways that platforms collect information on users and users disclose information about themselves and others. But crowd-level privacy norms can be complicated and refractory. Many crowds have privacy norms that are not embedded in the technical rules of the platform they share; the norms of appropriate flow are embodied only in ways that users habitually interact with each other.[22] Other crowds are dedicated to destroying privacy; a "human flesh search engine" is a crowd dedicated to investigating and bringing to light the perceived misdeeds of a specific target, and it requires a platform for members to share and collate the results of those investigations. Privacy laws must take account of the turbulent behavior at the platform-crowd interface.

- The competition-policy choice (such as it is) to have a few large platforms or many small platforms affects not just the welfare of users as individuals, but also the behavior of users in crowds.[23] Fragmented smaller platforms can nurture far more aggressive and toxic user communities, but also give their targets more defensible walls to pull back behind. Harassment simply plays out differently in a world of forums than in a world dominated by Facebook.

- The same is true in reverse. Legal interventions to promote user privacy, user data portability, and user safety all have effects on when and to what extent users come together in larger crowds. As such, they affect the competitive dynamics among platforms, making it easier or harder to platforms to compete on price or features. An empowered crowd can decide for itself when to leave one platform for another.

- Crowds also implicate the regulation of online content. Frameworks for free expression, content moderation, platform immunities, and common carriage typically begin with an individualistic model of users as independent actors, a model that the coherence of crowds subverts.[24] Reforms in these areas will need to account for how crowds use platforms for speech, how crowds will respond to those reforms, and whether crowds can exploit those reforms in unexpected ways.

Online policy interventions will need to take account of these linkages and effects.

Notes

1 *See* James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42 (2015), https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2620&context=facpub; Tarleton Gillespie, *Custodians of the Internet, Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (2021), https://yalebooks.yale.edu/9780300261431/custodians-of-the-internet.

2 *See, e.g.,* Alice E. Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age* (2013), https://yalebooks.yale.edu/9780300209389/status-update; Adrienne Massanari, *#Gamergate and the Fappening: How Reddit's Algorithm, Governance, and Culture Support*

Toxic Technocultures, 19 New Media & Soc'y 329 (2015), https://journals.sagepub.com/doi/abs/10.1177/1461444815608807; Anthony Nadler et al., *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech* (Data & Soc'y 2009), https://www.datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf; Rebecca Lewis, Alice E. Marwick & William Clyde Partin, *"We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube*, 65 Am. Behav. Scientist 735 (2021), https://journals.sagepub.com/doi/abs/10.1177/0002764221989781.

3 *See* Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 526, 545–48 (2022); Kate Crawford & Tarleton Gillespie, *What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Soc'y 410 (2014), https://papers.ssrn.com/abstract=2476464.

4 *See, e.g.,* J. Nathan Matias, *Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout*, 2016 Proc. CHI Conf. on Hum. Factors Comput. Sys. 1138, https://natematias.com/media/GoingDark-Matias-2016.pdf.

5 *See, e.g.,* Sulafa Zidani, *Represented Dreams: Subversive Expressions in Chinese Social Media as Alternative Symbolic Infrastructures*, *in* 4 Soc. Media + Soc'y No. 4 (2018), https://journals.sagepub.com/doi/full/10.1177/2056305118809512.

6 *See, e.g.,* Sarah Myers West, *Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms*, *in* 5 Media & Commc'n No. 3, at 28 (2017), https://www.cogitatiopress.com/mediaandcommunication/article/view/989.

7 *See* Dariusz Jemielniak, *Common Knowledge? An Ethnography of Wikipedia* (2014), https://www.sup.org/books/title/?id=24010; Amy S. Bruckman, *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge* (2022), https://www.cambridge.org/core/books/should-you-believe-wikipedia/F1797AA6843FEB206C2D7E418553C39C.

8 *See* Zeve Sanderson et al., *Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform*, *in* 2 Harv. Kennedy Sch. Misinformation Rev. No. 4 (2021), https://misinforeview.hks.harvard.edu/wp-content/uploads/2021/08/sanderson_twitter_trump_election_20210824.pdf.

9 *See* An Xiao Mina, *Memes to Movements: How the World's Most Viral Media Is Changing Social Protest and Power* (2019), https://www.penguinrandomhouse.com/books/567159/memes-to-movements-by-an-xiao-mina/.

10 *See, e.g.,* Shagun Jhaver et al., *The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action*, *in* 23 First Monday (2018), https://firstmonday.org/ojs/index.php/fm/article/view/8232/6644; Amanda Lenhart et al., *Online Harassment, Digital Abuse, and Cyberstalking in America* (Data & Soc'y Nov. 21, 2016), https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/; Alice E. Marwick, *Morally Motivated Networked Harassment as Normative Reinforcement*, *in* 7 Soc. Media & Soc'y (2021), https://journals.sagepub.com/doi/full/10.1177/20563051211021378.

11 *See* Stanford Internet Observatory, *Reply-Guys Go Hunting: An Investigation into a U.S. Astroturfing Operation on Facebook, Twitter, and Instagram* (Oct. 8, 2020), https://stacks.stanford.edu/file/druid:vh222ch4142/facebook-US-202009.pdf; Nicholas Confessore et al., *The Follower Factory*, N.Y. Times, Jan. 27, 2018, https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

12 *See* Deen Freelon et al., *Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation*, 40 Soc. Sci. Comput. Rev. 560 (2020), https://journals.sagepub.com/doi/abs/10.1177/0894439320914853; Paul Charon & Jean-Baptiste Jeangéne Vilmer, *Chinese Influence Operations: A Machiavellian Moment* ch. 9 (Oct. 2021), https://www.irsem.fr/report.html.

13 *Cf.* Salome Viljoen, *A Relational Theory of Data Governance*, 131 Yale L.J. 573 (2021), https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance.

14 *See* E. Gabriella Coleman, *Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle*, *in The Social Media Reader* 44 (Michael Mandiberg ed., 2012), http://media-study.com/resources/pdfs/socialmedia.pdf.

15 *See, e.g.,* Yiqing Hua et al., *Characterizing Alternative Monetization Strategies on YouTube*, *in* 6 Proc. ACM on Hum.-Comput. Interaction art. 283 (2022), https://dl.acm.org/doi/pdf/10.1145/3555174.

16 *See* Dhruv Aggarwal et al., *The Meme Stock Frenzy: Origins and Implications*, 96 S. Cal. L. Rev. 1387 (2024).

17 *See* David S. Evans et al., *Invisible Engines: How Software Platforms Drive Innovation and Transform Industries* (2006), https://mitpress.mit.edu/books/invisible-engines.

18 *See* Herbert Hovenkamp, *Antitrust Interoperability Remedies*, 123 Colum. L. Rev. F. 1 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4035879; Mike Masnick, *Protocols, Not Platforms: A Technological Approach to Free Speech* (Knight First Amend. Inst. at Columbia Univ. 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

19 *See* Chris Stokel-Walker, *Twitter May Have Lost More than a Million Users Since Elon Musk Took Over*, MIT Tech. Rev. (Nov. 3, 2022), https://www.technologyreview.com/2022/11/03/1062752/twitter-may-have-lost-more-than-a-million-users-since-elon-musk-took-over/.

20 *Compare* Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 Pepp. L. Rev. 9 (2009), https://digitalcommons.pepperdine.edu/plr/vol36/iss2/9, *with* Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. Rev. 713 (2023), https://www.bu.edu/bulawreview/files/2023/10/CITRON.pdf.

21 *See* Douek, *supra* note 3, at 545–46.

22 *See* David Auerbach, *Anonymity as Culture: Treatise*, Triple Canopy, Feb. 9, 2012, https://www.canopycanopycanopy.com/issues/15/contents/anonymity_as_culture__treatise; Gabriella Coleman, *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*

(2015), https://www.versobooks.com/books/2027-hacker-hoaxer-whistleblower-spy.

23 *See generally* Nikolas Guggenberger, *Essential Platforms*, 24 Stan. Tech. L. Rev. 237 (2020) [hereinafter Guggenberger, *Essential*], https://law.stanford.edu/wp-content/uploads/2021/05/publish_this_-_guggenberger_essential_platforms_eic.pdf.

24 *See generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2018), https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/; Eugene Volokh, *Treating Social Media Platforms like Common Carriers?*, 1 J. Free Speech L. 377 (2021), https://papers.ssrn.com/abstract=3913792.

Chapter 3

What Makes a Crowd?

Jessica L. Beyer

Every day, strangers find themselves talking about a wide range of topics, reading the same news article, and playing games together in common online spaces. Sometimes these strangers squabble and go on their way—but sometimes they form crowds. To form crowds, people often build a sense of community with other people in online spaces, sometimes then mobilizing collectively, working together to achieve a common collective goal.[1] For a crowd to exist, there must be a sense of "we" that supports and fuels group interaction and collective behavior.[2] Most of the time, when people return to an online space regularly a sense of community forms—or, if not a sense of community, a sense of place.[3] This feeling of community or co-presence with familiar others can be anywhere from deep to shallow, but it usually is built on a foundation of in-group identification.[4]

In-group identification comes from shared interests, terminology, behavior norms, jokes, and stories.[5] It can also come from shared beliefs that have either brought people to a space, such as a white

supremacist posting board;[6] shared beliefs that have developed within a space, such as an approach to getting babies to sleep; or shared interests, such as knitting.[7] These touchstones are co-created by the users in a space, with new users learning about them and being initiated into their evolving creation. Information is shared, educating casual observers and possibly recruiting them into the crowd.[8] The affordances of the internet mean that people are able to be collaborative in creating shared cultural products and often the memes or jokes or other materials that are produced are both tied to a particular community as well as referential to the broader political and social context of those in that space.[9]

Online spaces where individuals in the crowd can know each other, for instance, they have some distinct identity in that space including pseudonyms, are different from those in places with deeper anonymity, such as a place like 4chan. For example, in a place where people are known, there will be more individual ownership of goods and individually identified experts. In places with a high level of anonymity, collective production is more likely.[10]

One way that crowds develop is when a group is connected to a larger identity ecosystem. If the group is part of a larger identity ecosystem, such as a gaming community or a political group, the shared content ties people to that broader context, reinforcing and building a crowd. In-group jokes may reference history in the shared

In-group identification comes from shared interests, terminology, behavior norms, jokes, and stories.

game, stereotypes of opposing political groups, or hate speech.[11] Some communities, particularly those that are pursuing a particular agenda, will deploy collectively produced materials such as political memes and use networks and pathways that activists or commercial actors may use.[12] In these cases, political framing or references usually impact what is created explicitly. This type of crowd formation gains a lot of attention because it affects broad contestations for power, and within those contestations crowds become more visible. Sometimes, this crowd behavior may occur within a political framework such as an election, but sometimes people in a particular online space may present a shared set of values that are then articulated as having political significance by organized activists. This was the pattern with file-sharing sites. On these sites, people were sharing files online for a number of reasons, but activists, such as those running the file sharing site The Pirate Bay, reframed and rearticulated their behavior as having political meaning.[13]

Crowds can also form because of an information deficit or because of mistrust of authorities. Some online communities' reason for being is to provide information because community members feel there is ambiguity about what is true, often in situations that are high stakes to participants. In these communities, a feeling of camaraderie often develops, particularly because the information shared can be personal. One example of these kinds of spaces are online parenting forums. New parents are presented with an array of decisions that the popular press and online information sources frame as critical to a baby's short- and long-term well-being. These decisions are endless and are often framed in extreme but opposing terms. For instance, site participants may agree that feeding a baby the wrong foods will limit their intelligence, health, and happiness, but what constitutes the "right" food is hotly contested. Research illustrates that parents turn to the internet

for help navigating these questions.[14] Online forums that give parents information on vaccination are not always the organized anti-vaccine Facebook pages that Facebook has taken steps to ban.[15] Instead, they are often Facebook groups that are created and moderated by volunteers with content created by users. These users are asking questions in a crucial informational moment that is characterized by fear of the long-term consequences of decisions. The choice to vaccinate or not is an example of one of these crucial decisions. Because of the personal nature of the interactions in these spaces, trusted information networks may develop that then are the perfect transmission point for bad information or for new people to be exposed to broader disinformation networks.[16] These informational networks can also serve as a foundation for organized or parallel mobilization around an issue like vaccination.

Crowds also form through offline affinity or identity groups and their leaders shaping the way that people find their way to information online, which can include channeling people into particular online spaces or networks. For example, religious leaders may structure search terms and research methods in order to channel followers into informational eddies online, where they will find current events articulated to them in such a way as to give them a broader sense of group than their offline fellowship does.[17]

Crowds may draw upon these shared experiences, values, and identities to engage in collective action. For instance, Earl and Kimport[18] and Earl and Schussman[19] examine the ways in which collective action that most people would define as normatively "good" emerges from sites such as online fandoms, as well as others. Such sites draw on the affordances of the internet to facilitate a range of activism, including registering people to vote, asking people to sign petitions, organizing protests, and coordinating hacktivist action, such as DDOS attacks.

> Internet affordances create opportunities for crowds that most would define as normatively good, as well as those most would define as normatively bad.

Other crowds may engage in behavior that is normatively "bad" and often meant to cause harm. Organized harassment tactics such as doxing, brigading, crowd-level griefing, astroturfing, and DDOS attacks are examples. Crowds with any kind of agenda that are operating online learn from each other.[20]

However, collective goods sometimes also remain contained in a particular community and only resonate within that community, causing people to mobilize within the constructs of that space. For instance, on one site I observed, users would occasionally "riot" and flood the board with threads on a particular topic.[21] These riots were usually references to community jokes and contained, but sometimes riots would occur in reference to earlier riots. Mobilization like this can be disruptive to platforms needing to moderate interaction in a particular space, but they do not necessarily have broader implications.

It is easy to think of crowds as a problem to be solved because the affordances of online platforms and the internet amplify long standing human behaviors, both good and bad. Crowds are often a content moderation issue in that they can present a myriad of problems including whether the crowd's behavior norms are in line with those the platform would like to host; whether the crowd represents a political

position that is divisive and could reflect back on the platform itself; and whether the crowd is engaging in dangerous or violent behavior, such as doxing, or spreading bad information that could harm people, such as anti-vaccine rhetoric. However, although crowds can pose problems that range from annoying to dangerous for those managing interaction in online spaces, it is important to acknowledge their ambiguity. Internet affordances create opportunities for crowds that most would define as normatively good, as well as those most would define as normatively bad.

Notes

1 *See* Jessica L. Beyer, *Expect Us: Online Communities and Political Mobilization* (2014), https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199330751.001.0001/acprof-9780199330751; Gabriella Coleman, *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous* (2015), https://www.versobooks.com/books/2027-hacker-hoaxer-whistleblower-spy; Jennifer Earl & Katrina Kimport, *Digitally Enabled Social Change: Activism in the Internet Age* (2011).

2 *See* W. Lance Bennett & Alexandra Segerberg, *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*, 15 Info. Commc'n & Soc'y 739 (2012); Jessica L. Beyer, *Trolls and Hacktivists: Political Mobilization from Online Communities*, *in The Oxford Handbook of Digital Media Sociology* 417 (Deanna A. Rohlinger & Sarah Sobieraj eds., 2022); Francesca Polletta & James M. Jasper, *Collective Identity and Social Movements*, 27 Ann. Rev. Socio. 283 (2001).

3 *See* Beyer, *supra* note 1.

4 *See* Beyer, *supra* note 1; Coleman, *supra* note 1.

5 *See* Beyer, *supra* note 1; Jean Burgess, *"All Your Chocolate Rain Are Belong to Us"?: Viral Video, YouTube and the Dynamics of Participatory Culture*, *in Video Vortex Reader: Responses to YouTube* 101 (Geert Lovink

& Sabine Niederer eds., 2008); Coleman, *supra* note 1; Whitney Phillips, *This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture* (2015).

6 Jessie Daniels, *Cyber Racism: White Supremacy Online and the New Attack on Civil Rights* (2009); Whitney Phillips, Jessica L. Beyer & Gabriella Coleman, *Trolling Scholars Debunk the Idea that Alt-Right Shitposters Have Magic Powers* (Mar. 27, 2017), https://www.vice.com/en/article/z4k549/trolling-scholars-debunk-the-idea-that-the-alt-rights-trolls-have-magic-powers.

7 *See* Maura Kelly, *Knitting as a Feminist Project?*, 44 Women's Stud. Int'l F. 133 (2014).

8 *See* Beyer, *supra* note 1; Coleman, *supra* note 1; Deen Freelon, Charlton D. McIlwain & Meredith Clark, *Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice* (Ctr. for Media & Soc. Impact 2016), https://cmsimpact.org/resource/beyond-hashtags-ferguson-blacklivesmatter-online-struggle-offline-justice/.

9 *See* Daniels, *supra* note 6; Rebecca Lewis, Alice E. Marwick & William Clyde Partin, *"We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube*, 65 Am. Behav. Scientist 735 (2021), https://journals.sagepub.com/doi/abs/10.1177/0002764221989781; Alice Marwick & Rebecca Lewis, *Media Manipulation and Disinformation Online* (Data & Soc'y May 15, 2017), https://datasociety.net/library/media-manipulation-and-disinfo-online/; Phillips, *supra* note 5; Sarah Sobieraj, *Credible Threat: Attacks Against Women Online and the Future of Democracy* (2020).

10 *See* Beyer, *supra* note 1.

11 *See* Jessica L. Beyer, *Women's (Dis)embodied Engagement with Male-Dominated Online Communities*, *in Cyberfeminism 2.0*, at 153 (Radhika Gajjala & Yeon Ju Oh eds., 2012); Beyer, *supra* note 1; Mia Consalvo, *Confronting Toxic Gamer Culture: A Challenge for Feminist Game Studies Scholars*, 1 Ada: A J. Gender New Media & Tech. 1 (2012); Daniels, *supra* note 6; Kishonna L. Gray, *Deviant Bodies, Stigmatized Identities, and Racist Acts: Examining the Experiences of African-American Gamers in*

Xbox Live, 18 New Rev. Hypermedia & Multimedia 261 (2012); Kishonna L. Gray, *Intersecting Oppressions and Online Communities: Examining the Experiences of Women of Color in Xbox Live*, 15 Info. Commc'n & Soc'y 411 (2012); Kishonna L. Gray, Bertan Buyukozturk & Zachary G. Hill, *Blurring the Boundaries: Using Gamergate to Examine "Real" and Symbolic Violence against Women in Contemporary Gaming Culture*, *in* 11 Socio. Compass No. e12458 (2017); Lewis, Marwick & Partin, *supra* note 9; Alice E. Marwick & Robyn Caplan, *Drinking Male Tears: Language, the Manosphere, and Networked Harassment*, 18 Feminist Media Stud. 543 (2018), http://www.tiara.org/wp-content/uploads/2018/05/Marwick_Caplan_Drinking-male-tears-language-the-manosphere-and-networked-harassment.pdf; Marwick & Lewis, *supra* note 9; Adrienne Massanari, *Gamergate*, *in The International Encyclopedia of Gender, Media, and Communication* (K. Ross et al. ed., 2020); Sobieraj, *supra* note 9.

12 *See* Daniels, *supra* note 6; Emily K. Carian & Tagart Cain Sobotka, *Playing the Trump Card: Masculinity Threat and the U.S. 2016 Presidential Election*, 4 Socius 1 (2018), https://journals.sagepub.com/doi/full/10.1177/2378023117740699; Pierce Alexander Dignam & Deana A. Rohlinger, *Misogynistic Men Online: How the Red Pill Helped Elect Trump*, 44 Signs: J. Women Culture & Soc'y 589 (2019); Sobieraj, *supra* note 9.

13 *See* Beyer, *supra* note 1; Jessica L. Beyer & Fenwick McKelvey, *You Are Not Welcome Among Us: Pirates and the State*, 9 Int'l J. Commc'n 890 (2015); Fenwick McKelvey, *We Like Copies, Just Don't Let the Others Fool You: The Paradox of the Pirate Bay*, 16 Television & New Media 734 (2015).

14 *See* Jennifer D. Furkin, Mom to Mom: Online Breastfeeding Advice (2018) (unpublished doctoral dissertation), https://uknowledge.uky.edu/comm_etds/64/; Jodi Dworkin, Jessica Connell & Jennifer Doty, *A Literature Review of Parents' Online Behavior*, 7 Cyberpsychology: J. Psychosocial Rsch. on Cyberspace no. 2, art. 2 (2013), https://cyberpsychology.eu/article/view/4284/3329; Juyoung Jing, Jodi Dworkin & Heather Hessel, *Mothers' Use of Information and Communication Technologies for Information Seeking*, 18 Cyberpsychology Behav. & Soc. Networking 221 (2015); Lars Plantin & Kristian Daneback, *Parenthood, Information and Support on the Internet. A Literature Review of Research on Parents and*

Professionals Online, *in* 10 BMC Family Prac. No. 34 (2009), https://bmcprimcare.biomedcentral.com/articles/10.1186/1471-2296-10-34.

15 *See* Louise Matsakis, *Facebook Will Crack Down on Anti-Vaccine Content*, Wired (Mar. 7, 2019), https://www.wired.com/story/facebook-anti-vaccine-crack-down/.

16 *See* Dworkin, Connell & Doty, *supra* note 14; Sheera Frenkel, *She Warned of "Peer-to-Peer Misinformation." Congress Listened*, N.Y. Times, Nov. 13, 2017, at B1, https://www.nytimes.com/2017/11/12/technology/social-media-disinformation.html.

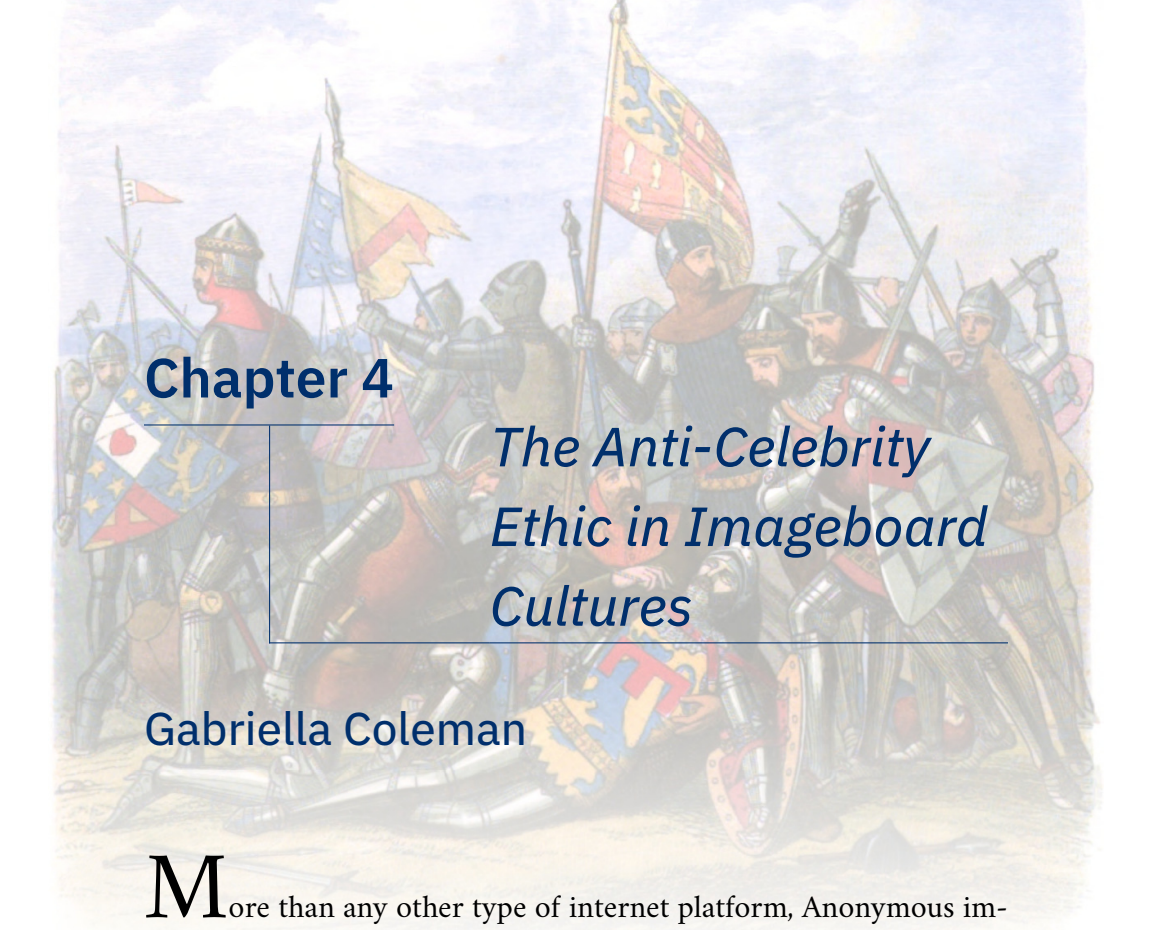
17 *See* Francesca Tripodi, *Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices* (Data & Soc'y May 16, 2018), https://datasociety.net/library/searching-for-alternative-facts/.

18 *See* Earl & Kimport, *supra* note 1.

19 *See* Jennifer Earl & Alan Schussman, *The New Site of Activism: Online Organizations, Movement Entrepreneurs, and the Changing Location of Social Movement Decision Making.*, 24 Rsch. Soc. Movements Conflicts & Change 155 (2002); Jennifer Earl & Alan Schussman, *Cease and Desist: Repression, Strategic Voting and the 2000 US Presidential Election*, 9 Mobilization 181 (2004).

20 *See* An Xiao Mina, *Memes to Movements: How the World's Most Viral Media Is Changing Social Protest and Power* (2019), https://www.penguinrandomhouse.com/books/567159/memes-to-movements-by-an-xiao-mina/.

21 *See* Beyer, *supra* note 1.



Chapter 4

The Anti-Celebrity Ethic in Imageboard Cultures

Gabriella Coleman

More than any other type of internet platform, Anonymous imageboards, like 4chan, 7chan, and 8chan, are the locus of what a group of law scholars has described as the "offensive internet." Not surprisingly, commentators and journalists often treat them as sites of pure lawlessness and chaos powered by a jacked-up anonymous mob. Essentially they are seen as places where anything goes. And yet those of us who have studied the boards for quite some time can see that despite free-wheeling attitudes (and yes, jaw-droppingly shocking and offensive content and dynamics), stable norms, collective action, and even full-fledged political formations have taken shape there. Indeed, what is remarkable is how these boards have influenced the formation of multiple, consequential, and relatively stable political expressions and currents. Whether it be the heyday of internet trolling circa the mid-2000s, the surprising hatching of Anonymous, a h/acktivist movement between 2008 and 2010, or the more recent flourishing of

the anonymous reactionary right and QAnon, these boards generate stable norms, collectives, and formations that sometimes have played outsized roles in political life outside of the boards themselves.

In this compendium, Jessica L. Beyer has addressed different facets of online crowd formation, including stranger-based online sociality, considering how such groups form, engage in collective action, collaborate, draw from and reassert dominant ideologies, and more. To further unpack one facet of the relationship between the individual and the collective within the context of these anonymous image boards, I will concentrate on a single example around stranger sociality. I will drill deep into the question of anonymity and specifically its ethical life, for how it influences and modulates the relationship between the individual and the collective in the context of imageboard culture.

I will begin by sketching the formation of this ethic or how it came to be on 4chan. Then I will briefly compare and contrast how it's operationalized and lived among the h/acktivists Anonymous and the anonymous far right, two stable—through quite distinct—formations to have come out of these boards. Here we will see how a robust anti-celebrity ethic formed that encouraged individuals to sublimate the self and identify with the group but also see how it shifted in subtle but important ways in the context of these two formations.

Back in the mid-2000s, a baseline commitment to anonymity was born on 4chan, thanks largely to technical features backed by social norms. One of 4chan's most distinctive sociotechnical features is anonymity: posts on the imageboard are attributed to "Anonymous," with the option to fill in a name field. But no one did that; instead, the community preferred the shared authorship and shunning of individual reputation, which eventually grew into a full-blown ethic, which persists today, though it has changed distinctly among distinct groups and formations that have come out of these boards.

A former Anonymous troll turned subsequent member of the Anonymous-led Anti-Scientology brigade explained the initial commitment to anonymous well: "The posts on 4chan have no names or any identifiable markers attached to them. The only thing you are able to judge a post by is its content and nothing else. This elimination of the persona, and by extension everything associated with it, such as leadership, representation, and status, is [became] the primary ideal of Anonymous."

I substituted *"is"* with *became* because initially, anonymity existed as a perfect example of what Raymond Williams defined as a "structure of feeling," or what we might approach as an infrastructure of feeling: "social experiences *in solution*, as distinct from other social semantic formations which have been precipitated and are more evidently and more immediately available." What this means is participants experienced anonymity more than they theorized it. Eventually, around 2005, board participants started to precipitate anonymity as a core ideal partly by using the collective name "Anonymous" to designate a collective, as they started to flesh out an explicit moral code. The naming of troll campaigns as indebted to "Anonymous" was part of this crystallization but a more explicit code that embodied various facets,

Participants came to identify with the ideal of anonymity and a general anti-celebrity ethic. But this sensibility would change.

idealized meritocracy and cast anonymity as a virtuous alternative to our celebrity-obsessed culture.

In this era (2005–2008), conversation, at times, gave way to collective action with trolling raids and campaigns. Some, but certainly not all, participants from 4chan and similar boards would coordinate, sometimes on invasion boards and chat rooms, swarming events that often led to doxing, harassment, prank calling, swatting or tormenting people online for the lolz, their own enjoyment (technically 4chan banned invasion boards, so Anonymous set up shop elsewhere too). Anonymous-style trolling was episodic, swarm-like, and prolific enough that, in 2007, Fox News anointed Anonymous as the "Internet hate machine." And even if it was often hateful, it had not yet hooked consistently into any broader political mobilizing on the left or right. However, the commitment to anonymity only grew more pronounced.

In the context of these raids, anonymity acted as a protective cover/shield for those individuals who banded together to unleash their campaigns of pranking, harassment and so on. But many board participants (many who also did not contribute to the raids) came to identify with the ideal of anonymity and a general anti-celebrity ethic. But this sensibility would change as some collectives broke away from the boards, and other new collectives took hold on the boards.

Six months after the Fox story, trolls used the moniker Anonymous to target the Church of Scientology—a campaign that morphed into a sincere crusade against the cult, protests that continued for nearly a decade under the Anonymous and Chanology banner. Crucial to this transformation was a video first designed as a troll that surprised both insiders and outsiders by prompting an earnest debate about protesting the Church. Critics of the Church also swooped in to urge these trolls to join their cause. They did. Anonymous moved forward with

an experimental but successful protest on February 10, 2008, carried out in 127 cities worldwide with over 7000 people showing up. They adopted the Guy Fawkes mask to insulate themselves from harm—a pop culture icon symbolizing the movement.

Anonymous thus started to become political out of a practice of apolitical (but often terrifying) trolling. Still, between 2008 and 2010, Anonymous existed in a transitional, brackish state: the name itself was contested. Trolls *and* activists coordinated campaigns with the same name and symbols. Indeed many of the 4chan troll types were livid that do-gooders had adopted the name and sought to reclaim the name Anonymous by dragging it through the mud, using terrifying campaigns like swarming epilepsy boards with flashing gifs to induce seizures.

After 2010, when quarters of Anonymous became more firmly rooted in activism and hacktivism and fully pivoted away from the boards and coordinated on chat rooms, Anonymous' ethical commitment to anonymity underwent a significant metamorphosis, a point made by the same Anonymous activist I quoted previously in this piece: "It became more nuanced . . . incarnating into the desire for leaderlessness and high democracy." I've documented this amply in my writing elsewhere. Still, I'd like to offer a small token example from Twitter of what this looks like:

> FemAnonFatal is a Collective • NOT an individual movement NOT a place for self-promotion NOT a place for HATE BUT a place for SISTERHOOD It Is A place to Nurture Revolution Read Our Manifesto . . . • You Should Have Expected Us • #FemAnonFatal[1]

"This desire for leaderlessness and high democracy," with a few exceptions like these tweets, was submerged from public view; it was largely visible to those like myself and a few others interact-

ing closely with Anonymous. More so, the norm was not only idealized but socially enforced. In chat rooms where Anonymous launched its operations, most chose to be pseudonymous and thus interacted with stable nicknames. Reputation, based on what you said and did, accrued. Soft leaders and more visible players emerged. And yet when people were seen to pine for attention, or especially when they acted non-anonymously, they were routinely cut down or kicked out.

In 2011, as Anonymous was scheming on internet relay chat rooms and scoring victories in the media after major hacks, back on 4chan, a new microculture was talking itself into existence. While politically incorrect and racist speech had been present from 4chan's inception, a board called /pol/ was created for "discussion" with a politically incorrect bent. A growing collection of users were engaging in discussion of news and ideology, bringing in interpretations that were so reactionary and racist that they would have been impossible to sustain in non-anonymized settings. This fact was not lost on other dark corners of the internet, and /pol/ became a real magnet for those active in the thriving neo-reactionary blogging communities, white nationalist message boards like Stormfront, and offline far-right organizations like the American Third Position Party.

Slowly, /pol/ grew. Moderators of other 4chan boards continued to direct their most racist and offensive users to what many moderators initially saw as a "containment board." While the term "alt-right" started to be bandied about at this point (that is, between 2011 and 2014), these anonymous posters on /pol/ didn't have a label or even a consistent, clear message, even as they used disinformation tactics that the reactionary or red-pilled right would eventually adopt more deliberately, as that formation became more stable, and especially after events like GamerGate.

Indeed, starting around 2014 but accelerating in 2016, the reactionary or far right—some of it anonymous, much of it not—deployed everything the internet had to offer—image boards, Twitter, Reddit, Facebook, Youtube, chat rooms—to establish itself as a political force. Composed of a loose coalition, it includes hard-core internet trolls, white nationalists and Nazis, men's rights activists, some libertarians, and many anonymous actors. They are united by a demonization of those they've dubbed SJWs or "Social Justice Warriors," figures they perceive to be forcing a "woke agenda" on the world. Through red pilling, they seek to awaken people to the harms of woke politics. To be red-pilled means to be awakened to a previously hidden truth; this can mean: "realizing" that the left is totalitarian, or that feminism is responsible for the decline of Western values, or that the MSM (mainstream media) is a mouthpiece of the liberal elite. While many far-right groups like the Proud Boys are not anonymous, anonymous image boards were key for this site of "metapolitics"—the idea that profound political change comes not only from the vote or policy but through culture itself.

The little "a" anonymous far right on /pol/ and other boards, in many ways, continued to embrace and uphold the anti-celebrity, anti-leader ethic first cultivated in the mid-2000s, which left me surprised (and disappointed), as I assumed it was a value more aligned with progressive politics and would wither away. Like Anonymous, among the anonymous red-pilled right, public figures associated with their cause or ideologies were maligned as "e-celebs" and constantly torn apart, undermined, and insulted from every possible angle. Here is a typical comment that captures their critiques of the e-celebs as parasites and frauds, making a buck while they are the authentic, true source of ideas and manipulations. "It should be clear by now that these frauds do nothing but try to sell your ideas back to

you . . . Nothing they promote is their own creation (exception themselves)."

In contrast to Anonymous, where such figures were in fact, sidelined, in the case of the far right, many of these figures were still seen as useful—very useful—to their cause—a trope that was oft-repeated and normalized through hefty rationalization in very long threads as to why they put up with these otherwise disdained figures. In their own words: "pol leads the right. Milo and all the other e-celebs deliver the cause to normies." They were instrumentalized and put up with as a necessary evil to spread the overall gist and message to as many people as possible.

This detail is small but important. In both currents, Anonymous and the anonymous far right, the anti-celebrity ethic is meant to and does support in-group solidarity. They both routinely attacked individualistic, hypocritical, or egotistical behavior. But for the far right, the imperative for recruitment, mainstreaming their dangerous ideas, and reaching new actors superseded all else. Whereas in Anonymous, there was no such cost-benefit calculus, the norm generally carried out to its full conclusion, a prefigurative mechanism for modulating and lessening power and hitched rather closely to other visions of social good: equality for all types of people. This type of message is impossible to inculcate in the far right, who otherwise embrace a social vision where hierarchy, power, and authority are naturalized as worthy.

Notes

1 *See* Gabriella Coleman, *Reconsidering Anonymity in the Age of Narcissism*, 54 McSweeney's 195, 212 (2018).

Chapter 5

Fandom as a Wellspring and Leading Indicator

Rebecca Tushnet

Media fandom—fandom organized around particular cultural artifacts, involving the creation of fan fiction, fan art, fan video, and other fannish artifacts—has been the source of many significant cultural innovations online, for good and ill. Many people have suggested that Gamergate, organized ostensibly around video games and by video game fans, provided a model for subsequent right-wing online movements.

More generally, fan behaviors, including organizing online campaigns, now provide a model for political organizing as well as a source of community that itself can aid in forming a political consciousness, as Mimi Ito explained in 2009.[1] Ito recounts protests against American beef imports, significant enough that they threatened the entire South Korean government and led to the firing of the cabinet, along with an apology from the president for moving too fast without consulting the

> Participating in fandom is pleasurable. This is the source of its power and also its danger.

public. The protesters numbered over a million, an estimated 60–70% of which were teens, mostly teenage girls. Many were fans of a boy band, and they used their fandom as a source of organizational power. As Ito concludes, "you should never underestimate the power of peer-to-peer social communication and the bonding force of popular culture. Although so much of what kids are doing online may look trivial and frivolous, what they are doing is building the capacity to connect, to communicate, and ultimately, to mobilize."

Fan creativity also powerfully illustrates the importance of a varied, participatory online ecosystem: Although most individual fanworks are not very good, some are brilliant, and they wouldn't exist without the enthusiasm and encouragement of other fans. No creative field operates with only geniuses; just as Shakespeare emerged from a welter of justifiably forgotten playwrights, we still need a lot of dross to get the gold. Attachment to a fannish object—whether a boy band, a movie, a book series, or something else—encourages people, especially young women, to see themselves as potential creators, with significant benefits both for skills such as language proficiency and video editing as well as for self-confidence and self-understanding.[2]

Participating in fandom generates lots of activity because it's pleasurable. This is the source of its power and also its danger: commitment to fandom and fellow fans can be a barrier to rejecting misinformation, and can encourage negative coordinated action, including

harassment and manipulation of online results, even as it can also encourage positive coordinated action such as disseminating truthful information, making charitable contributions, and engaging in advocacy for marginalized groups.[3] Fandom has encouraged underrepresented groups to participate more openly in public life, but it is no utopia, especially for fans of color who may find themselves excluded or harassed.[4] Even authoritarian governments have struggled to contain fans' energies.[5]

Because of the varied activities and passions they inspire, fandom organizations and experiences provide lessons for content moderation more generally.

First, every content moderation tool is also a tool of abuse. Mechanisms for reporting bad content will also be used to harass legitimate creators. For example, fans of one celebrity or a fictional character may target fans of another celebrity or character who they perceive as competing with their favorite. Anti-impersonation rules will also be used to target critics and people with good reason to conceal their identities, as well as people who are just playing with identity and not deceiving anyone. Blocking will also be used to suppress dissent, as when politicians or official entities delete negative comments and block citizens who disagree with them. Historically, regulators have been interested in making content moderation tools readily available, for example allowing copyright claimants to send takedown notices, but they have not built in effective means for preventing abuse of those tools, which

At scale, individual procedure can be the enemy of overall justice.

should include room for platforms to reject low-quality reports outright.

Second, every tool that is useful to good-faith users is also a tool of abuse. The ability to mention or tag other users, the ability to link to other users' content, the ability to comment on other users' content, and every other valuable feature that links users together will also be misused to harm people. Fans benefit from being able to comment on each others' works and connect in other ways, but they can also use these tools to harass and urge others to harass a target. Both platform designers and regulatory designers must remain aware of this inevitable dual-use capability rather than presuming that users will in general make only one kind of use of any given feature.

Third, abusive conduct occurs cross-platform and one platform may have minimal or no insight into why the part of the conduct that occurs on its platform is or contributes to a problem. In fandom, a victim of harassment may be targeted across multiple platforms so that each platform can't see the full scope of the problem. Cross-platform abuse also occurs with counterfeiting, where advertisements on one site lead to a legitimate-looking product on another site. The status of such "hidden" counterfeits are invisible to the second site that is actually facilitating the sales. Likewise, solicitations to share child sexual abuse material (CSAM) through coded invitations can benefit from leveraging different platforms. The infinite adaptability of actors who desire to behave badly means that platforms will always be playing catch-up and fighting the last war.

Finally, at scale, individual procedure can be the enemy of overall justice. Every platform faces some kind of resource constraints. Fan-run platforms, such as the nonprofit Archive of Our Own, are run by volunteers with limited resources and time, but even when moderators are paid, decisions are inevitably required about what to priori-

tize. Where there are a hundred thousand—or more—decisions to be made, a focus on giving maximum due process to individual user appeals can detract from the overall health of the system by discouraging moderation of squeaky wheels and absorbing resources that could be used in more proactive monitoring. This is especially significant because people who already feel empowered—usually men from dominant groups—are more likely to appeal adverse decisions, worsening systemic bias. It is possible to drown in procedure without achieving substantive justice.

Notes

1 *See* Mimi Ito, Keynote Address at the 51st NFAIS Annual Conference: Media Literacy and Social Action in a Post-Pokemon World (Feb. 24, 2009), http://www.itofisher.com/mito/publications/media_literacy.html; *see also* Ashley Hinck, *Politics for the Love of Fandom: Fan-Based Citizenship in the Digital Age* (2019); Henry Jenkins, *Fans, Bloggers, and Gamers: Exploring Participatory Culture* (2006); Kaitlyn Tiffany, *Everything I Need I Get from You: How Fangirls Created the Internet as We Know It* (2022); Liesbet van Zoonen, *Entertaining the Citizen: When Politics and Popular Culture Converge* (2004).

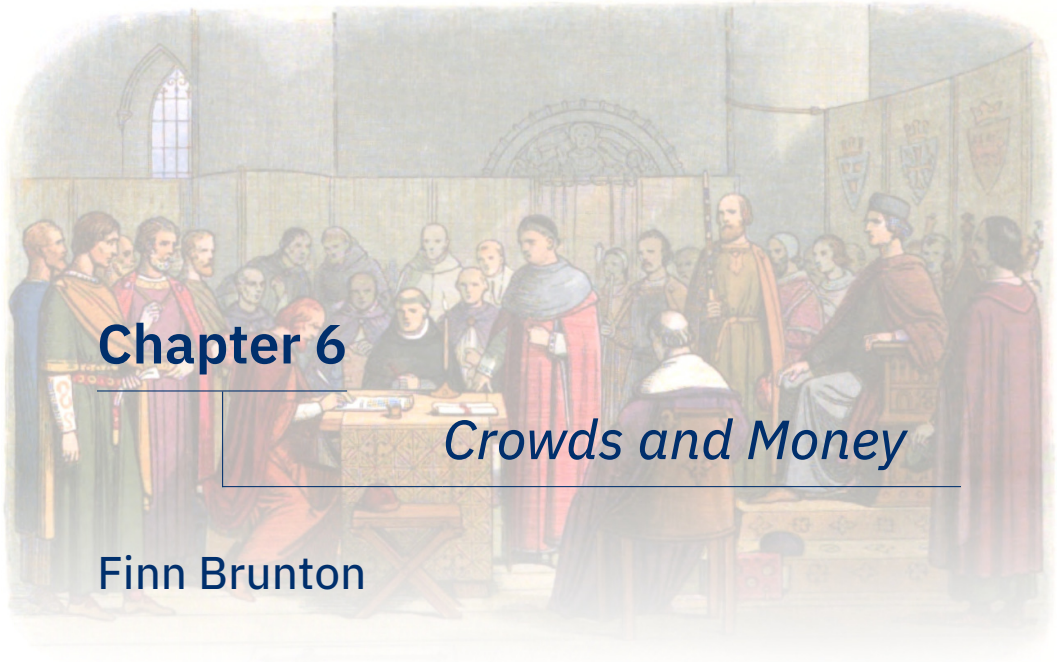
2 *See, e.g.,* Betsy Rosenblatt & Rebecca Tushnet, *Transformative Works: Young Women's Voices on Fandom and Fair Use*, https://books.openedition.org/uop/524, *in Egirls, Ecitizens: Putting Technology Theory, Policy And Education Into Dialogue With Girls' And Young Women's Voices* 385 (Jane Bailey & Valerie Steeves eds., 2015); Comments of the Organization for Transformative Works, Department of Commerce Green Paper, Copyright Policy, Creativity, and Innovation in the Digital Economy, 78 Fed. Reg. 61337 (Nat'l Telecomms. & Info. Admin. Nov. 13, 2013) (No. 130927852-3852-01).

3 *See, e.g.,* Taylor Behnke, *The Blackpink Fans Looking to Outfox YouTube*, Vulture (Sept. 19, 2022), https://www.vulture.com/2022/09/blackpink-fans-born-pink-youtube-streams.html; Jin Ha Lee et al.,

Community-Based Strategies for Combating Misinformation: Learning from a Popular Culture Fandom, Misinformation Rev. (Sept. 28, 2022), https://misinforeview.hks.harvard.edu/article/community-based-strategies-for-combating-misinformation-learning-from-a-popular-culture-fandom/; Elizabeth Williamson, *"Prove to the World You've Lost Your Son"*, Slate (June 13, 2022), https://slate.com/human-interest/2022/06/shooting-school-texas-uvalde-sandy-hook-conspiracy.html.

4 *See, e.g.,* Rukmini Pande, *Squee from the Margins: Fandom and Race* (2018).

5 *See* Aja Romano, *The Chinese Government's Unlikeliest Stand-off Is With . . . Fandom*, Vox (Oct. 17, 2022), https://www.vox.com/culture/23404571/china-vs-fandom-danmei-censorship-qinglang-social-media.



Chapter 6

Crowds and Money

Finn Brunton

> [They are] two perfectly insignificant and incapable individuals, whose existence is only rendered possible through the high organization of civilized crowds. Few men realize that their life, the very essence of their character, their capabilities and their audacities, are only the expression of their belief in the safety of their surroundings. The courage, the composure, the confidence; the emotions and principles; every great and every insignificant thought belongs not to the individual but to the crowd: to the crowd that believes blindly in the irresistible force of its institutions and of its morals, in the power of the police and of its opinion.
>
> —Joseph Conrad, *An Outpost of Progress* (1896)

Markets Are Crowds

Not all crowds are markets, but all markets are crowds. A religious gathering, a house party, a book club, or the Storming of the Bastille: crowds, but not markets. A market without a crowd, though, is a one-

hand-clapping thought experiment. A transaction or exchange takes two, and a market takes many. A market without a crowd is like a language spoken by only one person.

The crowd that makes up a market can be as varied in its form as any crowd in general.

Constituents A market, as a specialized type of crowd, can be all sellers and a single buyer, like contractors bidding for a job. It can be all buyers and a single seller, as with an auction. It can have a clear division between those roles, like a trade show with fees and badges for those allowed to vend. Or the two parties can be more fluid: in a stock market or the rare book business most agents are buyer and seller at once.

The constituents of a digital market-crowd are not necessarily human. In order of sophistication, they can include scripts and bots; recommendation, pricing, and trading algorithms; and now potentially more advanced forms of machine learning and artificial intelligence.

Space and time A market-crowd can, of course, be local and immediate: the ancient Athenian *agora*, or a North African *souk*, or a flea market in a parking lot. It can be remote and delayed: buyers and sellers across the geography of the Silk Road, for instance. It can be local and delayed, as with "silent trading" throughout the ancient African world (one group deposits goods in a place and withdraws; the other party leaves what they will exchange, and withdraws; the first group returns and either takes the exchange goods, closing the deal, or leaves them, continuing the negotiation), or newspaper classifieds. Or, finally, remote and immediate.

This last category has transformed the nature of the market-crowd in the last century and a half. This development began with the introduction of the stock ticker (first installed in 1867). This realtime feed

of market-crowd sentiment is the prelude to transactions executed by remote, immediate communication. The remote crowd, both delayed or immediate, is the primary market of social and platform media.

Rules and structure The market-crowd is bound by rules. The rules can be formal, like contracts, licenses, and requirements under which one operates to sell a house or buy a car, or the EULA and TOS of an e-commerce platform. The rules can be social, informal, and implicit: consider the many cultures in which haggling is expected, and the stated price just a starting point, or one of Erving Goffman's famous "breaching experiments"—tests of tacit social norms—in which his graduate students shopped in a grocery store by taking items from the carts and baskets of other shoppers in the aisles.

There many other ways to characterize various market-crowds: deliberate or incidental (like commuters on a subway platform); porous or impermeable; how they hail or signal new members, and how they communicate within themselves. One commonality among them all, in the language of investment analysis, is the problem of *sentiment* or *spirit*.

And Markets (Which Are Crowds) Have Some Interesting Properties

A market-crowd has feelings and moods, and its moods are of a peculiar kind: self-referential and self-reflexive. A key determinant of a market's activity is how the market feels about itself, and how it feels about how it feels about itself. Any participant in the market crowd must to some extent think about how they feel about the market, and part of that thinking is their estimate of how other participants also feel. A run, a panic, a bubble: all are to some extent, depending on the

specific context, dependent not just on how the market reacts but on how it understands its own reaction.

The most obvious examples of this are in financial markets (bull, bear, boom, bust), but all market-crowds are faced with this problem of sentiment and its management, whether panic buying toilet paper or driving up the price of baseball cards. Indeed, we accept money itself in the belief that it will be accepted from us; while each exchange of money may be local and immediate, it is always a transaction in the context of a great crowd who act as the ultimate guarantors of the value that moves through the exchange. (Recent cryptocurrency crises provide perfect illustrations of what happens when that distributed crowd's spirit fails and its sentiment turns.)

In a market crowd, therefore, fake causes can have real effects. Rumors and lies can be understood as such, without being any less actionable: the question is not whether you believe it, but whether you believe other people believe it, and how you behave as a consequence. In Lichtenberg's aphorism, a king can command on pain of death that everyone treat an ordinary stone as a diamond. Eli Lilly's stock took a dive after a Twitter post ("we are excited to announce insulin is free now") by a fake Eli Lilly account: did people sell because they believed it, or because they were concerned that other people would believe it, or because they were reacting automatically to market movements triggered by belief or belief in belief?

Facilitating New Crowds Entails Facilitating New Markets

I was part of the market-crowd at a Trader Joe's grocery store where the shelves were noticeably empty. Was the truck delayed? Yes, said my cashier, but did I know what the real problem was? TikTok, he said. He was in charge of the protein bar section. Sometimes a protein bar sells

out as fast as it arrives, chronically out of stock; he orders more, then demand drops back down to normal and he has a glut taking up space. After the fact, asking around, he learns that the bar has been recommended by one or another TikTok fitness influencer: they have convened a social media market-crowd that sweeps through the physical retail space.

Creating new crowds creates new markets. This point may seem trivial in the context of social media platforms, but consider it more broadly. The crowd gathered by TikTok is not just a market for advertisers: it hosts some uncounted number of other markets, one of which clears the shelves of protein bars at a California Trader Joe's. The crowd gathered on Instagram is not just a market for Meta, or for personal brand building influencers: Insta accounts sell halal lamb in Dubai, Japanese manga in Djakarta, and cosmetics in Singapore, with a WhatsApp phone number in the bio for the transaction and the goods delivered by scooter.

Telegram facilitates crowds of many kinds, from Russian and Iranian dissidents to American and Canadian QAnon conspiracists and antivaxxers. These crowds can be markets in themselves—QAnon culture is rife with scams of all kinds—but there are many other more

> In a market crowd, fake causes can have real effects. The question is not whether you believe it, but whether you believe other people believe it.

explicit market-crowds, especially for retail drug dealing. Twitter and Reddit are at once social media platforms and, again, a site of many market-crowds, most notably for this essay those in the cryptocurrency business. Reddit and Twitter did not directly facilitate a marketplace in cryptos, any more than TikTok had a direct marketplace in protein bars; rather, they enabled the hailing and convening of new crowds, with their laser-eye avatars, in-joke usernames, and jargon-dense bios, who could boost prices, make predictions, and otherwise keep the faith. In some cases their posts could move other markets.

Which Enables Our Current Situation: Augmented Manipulation

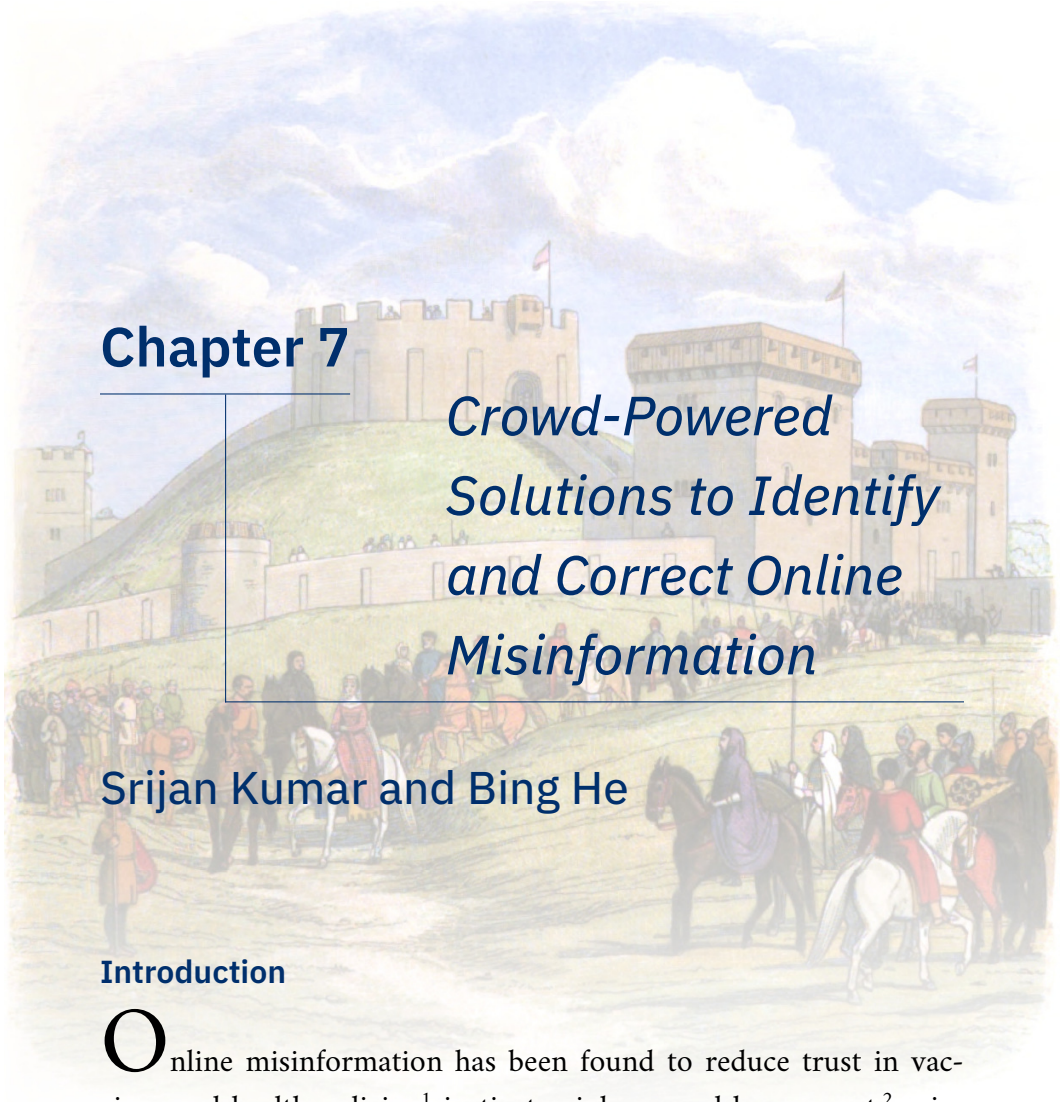
Of course, many of the posters in the cryptocurrency market-crowd were not human. They were social media bots, manipulating trends, boosting particular messages, and in some cases communicating directly with human participants. The mediated, remote, digital market-crowd can be characterized and distinguished from other types of market-crowds by its susceptibility to manipulation by the activity of non-human participants: *augmented manipulation*.

Many of the obvious examples of this are crude, blunt-edged instruments: the "snipers" that always win auctions with a one-penny bid at the last second, or the armies of dummy accounts that sweep up everything from concert tickets to streetwear drops for scalpers to resell. However, the takeaway I would like to leave with the reader is more complex in two ways, and the situation in cryptocurrency offers a good example for both.

The first takeaway is that a part-automated market-crowd is not much different to one without automation. Non-human market signals and actions are often indistinguishable from real market

activity—but wait: "real" market activity in what sense? We sometimes use "real" to mean the actions of authentic, independent humans, not automated bots. We also use "real" in this sense to mean true signals, rather than lies, rumors, and mendacious claims. But the market-crowd undercuts both of those uses of language. Many kinds of non-human actors engage in real market activity: following instructions to buy and sell, reacting to keywords in the news, looking for statistical patterns and responding to them. Their activity is not in any way different from much of the human activity in the crowd of which they are a part. And the humans often seem no less automated, reacting in ways that, if not rational, are still largely predictable. The same problem applies to their sometimes mendacious signaling and market activity: humans are just as prone as crews of bot accounts to circulate media and make purchases that can, they hope, set off a run on an asset they've shorted, or inflate the value of an asset they hold. At least the bots, as they auto-retweet false promises and generate the illusion of interest in a crypto property, do not release multi-hour podcasts and YouTube videos as part of their scam.

The second takeaway is that the fact of the augmented manipulation of the market-crowd has been incorporated into the crowd's theory of itself. Price fixing, pump-and-dump, managed bubbles and panics and the like are open secrets if they are secrets at all. The barons—crypto whales, financial brand managers, traders who hedge market sentiment—have their thumb on the scale, but that's no secret for the mob at this point. Any person in the crowd can follow the movement of cryptos on their various ledgers, and can learn to spot bots on their social networks and even to work with them. Their theory of the mind of the market-crowd can now assume and incorporate the fact of its augmented manipulation.



Chapter 7

Crowd-Powered Solutions to Identify and Correct Online Misinformation

Srijan Kumar and Bing He

Introduction

Online misinformation has been found to reduce trust in vaccines and health policies,[1] instigate violence and harassment,[2] raise skepticism about democratic processes and elections,[3] escalate polarization,[4] and detrimentally impact well-being.[5] Most people consume information and news from social media,[6] which has often become "ground-zero" for health misinformation and where misinformation tends to propagate faster and wider than truthful content.[7] COVID-19 vaccine misinformation, such as unsubstantiated claims about the vaccine leading to infertility, containing microchips, or altering DNA and genes, has spurred vaccine hesitancy, lowered vaccine uptake, and prolonged the pandemic. Furthermore, misinforma-

tion has had destructive effects on individuals. For instance, misinformation asserting that Bill Gates created COVID-19 vaccines for population control led to public distrust and verbal aggression.[8] Similarly, misinformation and disinformation about elections have reduced trust in democratic processes and institutions.[9] Racial misinformation has led to physical violence and online harassment against racial minorities, including Asians and Black communities.[10] Thus, it is critical to curb the spread of online misinformation.[11] Here, we use a broad definition of misinformation which includes falsehoods, inaccuracies, rumors, decontextualized truths, or misleading leaps of logic.[12]

Detection of misinformation has been studied extensively in recent years using machine learning methods, professional fact-checkers, and crowds.[13] A great deal of effort has gone in developing *machine learning models* for detecting misinformation[14] has relied on the post's content,[15] poster's attributes,[16] social network,[17] time,[18] and propagation features.[19] Many such machine learning models have been deployed on web and social media platforms that serve as automated filters to remove dangerous content.

On the other hand, *professional fact-checkers* provide fact check labels to questionable claims. They write fact-checking articles detailing the reasoning behind their fact-check label determination along with the evidence that is available to back up the label. For example, Snopes[20] provides fact check labels ranging from true and mostly true to mostly false and false. Many machine learning methods rely on these fact-check labels to train the detection models. While fact-checking has high precision, it has low throughput due to the primarily manual process of finding relevant evidence, digging into it, and providing labels. It typically takes a few hours to a few days for the first fact check to be released after a new questionable claim appears.

Despite these solutions, the misinformation infodemic[21] is still growing alarmingly. Reasons include that the burden of flagging misinformation is bottlenecked by the limited number of fact-checkers, time taken to generate the labels is high, fact-checks only cover a small number of viral claims, and most automated machine learning models depend on fact-check labels.[22] Automated machine learning models are slow to react to the changing information ecosystem and can be manipulated by adversaries.[23] There also exist no structured methods to counter misinformation after it has been identified[24] as fact-checkers do not actively engage with misinformation spreaders. All of this makes the existing solutions *reactive* to misinformation, i.e., detection and mitigation is done after misinformation has already spread.

Complementing the above two approaches, leveraging *crowds* provides a promising solution to mitigate misinformation via an approach that can be scalable, proactive, and effective. Crowds, i.e., non-expert social media users, act as eyes-on-the-ground who proactively question and counter misinformation.[25] Crowds can flag emerging misinformation, which makes it feasible to identify them early before they get attention. They complement fact checkers who only verify a handful of stories after they have gone viral.[26] Crowd engagement with (mis)information generates textual and non-textual signals that can improve existing machine learning-based (mis)information classification models.[27] Importantly, crowds actively also actively engage in the activity of countering misinformation, for example, by replying to misinformation posts with the correct information. In fact, our recent work has shown that 96% of counter-misinformation responses are made by ordinary users, while professionals account for the rest.[28] This is called *social correction* and has been shown to be as effective as professional correction,[29] curbs misinformation spread,[30] and works across topics,[31] platforms and demographics.[32]

This article discusses the state of the current research in computer science and social science to give a comprehensive overview of how crowds detect and correct online misinformation.

Misinformation Detection and Flagging with Crowds

This section discusses some research on the use of crowd knowledge to detect misinformation, via surveys and via machine learning algorithms. Next, Twitter's Birdwatch is discussed, which is a platform where crowds can flag misinformation.

Misinformation Flagging with Crowds Recent research has analyzed the role that crowds play in flagging misinformation. Crowd-generated misinformation flagging complements fact-check generation and correction by professionals, both of which have been studied extensively.[33]

Crowds can flag and annotate online misinformation at scale and these annotations are shown to be accurate.[34] In particular, after the majority voting or aggregation of individual scores,[35] the rating of crowds can be comparable to or outperform that of professionals. Even if some annotations from laypeople are noisy and ineffective,[36] a high level of annotation agreement among crowd workers assures their ability to identify rumors.[37] The decent annotation performance still holds even in highly politicized contexts.[38] Moreover, regardless of the ideological differences and lack of familiarity with many news outlets, Pennycook et al.[39] found the trust ratings of crowds are successful at differentiating mainstream media outlets from hyperpartisan and fake news websites, and the ratings were highly correlated with those of the professional fact-checkers. This further confirms the fact-checking capability of the crowds.

When crowds engage with (mis)information online, it generates a large volume of semantically rich content that supports, questions, expresses concern and disbelief, and counters misinformation.[40] These provide complementary and useful signals (reply text, like, re/share, etc.)[41] to traditional content and social network features.[42] These have been used to improve the performance of misinformation detection machine learning models. Thus, crowd-generated flagging of and engagement with online content serves as important and useful signal for early identification of misinformation.

Birdwatch (a.k.a. Community Notes)—A Platform for Crowd-based Misinformation Flagging The use of crowd-generated inputs to combat misinformation is gaining mainstream attention. One popular instance is Twitter's Birdwatch (now called Community Notes) that facilitates misinformation detection by crowds. On the platform, users can report suspicious and/or misleading tweets, as well as annotate tweets reported by others. Many have investigated this kind of crowd-generated flags[43] to derive different patterns. For example, Allen et al.[44] examined the influence of political partisanship during the crowds' annotation by analyzing existing data from the Birdwatch platform. They found users are more likely to (1) give negative annotations of tweets from counter-partisan users, and (2) flag annotations from counter-partisan users as unhelpful. This highlights the primary issue of partisanship in crowd-initiated labels.

Besides, recent studies show,[45] crowds actively identify tweets that they believe are misleading and provide contextual notes for debunking. They actually have different levels of debunking capability. This shows a need to improve the annotation skills of ordinary users.

Our recent research has, however, shown that such crowd-based flagging systems need to be designed with care and caution. Through

simulations, we showed that Birdwatch can be manipulated by motivated bad actors.[46] Even simple manipulation that adds "true" ratings to misinformation posts and "false" ratings to true posts can change the classification and ranking of posts, leading to misinformation posts being categorized as true and displayed to other users. The manipulation can work when bad actors leverage multiple accounts (either by controlling them directly via hacking or purchase, or by coordination across multiple motivated accounts), which is quite typical on social platforms. Thus, whenever crowds are involved, there is a need to create a reputation system to distinguish reliable crowd members from unreliable ones. Our research proposed one such reputation system and showed that it is less manipulatable by bad actors, the details of which can be found in the study.[47]

Misinformation Correction with Crowds

This section discusses research on misinformation correction by crowds, social correction theory, prevalence and patterns of social correction on social media platforms, and methods to empower better correction.

Social Correction Studies have shown remarkable effectiveness of social correction, i.e., countering of misinformation claims by the crowd, by conducting experiments via interviews,[48] surveys,[49] and in-lab experiments.[50] Social correction has proven to be as effective as professional correction,[51] curbs misinformation spread,[52] and works across topics,[53] platforms and demographics.[54] Notably, users' polite and evidenced responses that refute misinformation are shown to effectively counter misinformation and reduce the belief in misinformation.[55] Corrections work[56] without causing an increase in misperception (i.e., the backfire effect has not been replicated).[57] Users correct

others, typically friends,[58] owing to a sense of social duty,[59] anger, or guilt.[60] While corrections are not expected to convince everyone (e.g., users with extreme stance), they are most effective in reducing the misperceptions of misinformation consumers.[61]

Thus, empowering users to effectively correct misinformation promises a scalable solution towards information integrity.[62] Recent researches provide considerable evidence that correction by crowds is effective in countering misinformation and in mitigating the spread of misinformation.[63] Considering the limited capability of professional fact-checkers, the larger number of ordinary users and their efforts in social correction show great potential for a scalable solution to countering misinformation. Such a solution is independent of, but complements, the efforts of social media platforms to detect misinformation via the crowd, e.g., Twitter Birdwatch.[64]

Social Correction and Fact-Checking of Misinformation by Crowds on Social Media Platforms Crowds engage in both fact-checking and social correction on social media platforms. Fact-checking and social correction are slightly different, where fact-checking is done using a URL reference to a fact-checking website such as Snopes, while social correction can be done with or without providing a fact-checking URL. Thus, social correction encapsulates fact-checking.

Our recent works[65] have analyzed how users engage in social correction in-the-wild on Twitter, Facebook, and Reddit. In Micallef et al. (2020),[66] we collected 8 million COVID-19 tweets, hand-annotated tweets as misinformation or counter-misinformation, and built a language-based detection model using BERT to categorize the remaining tweets. Notably, our work looked at the linguistic properties to identify misinformation corrections, rather than just looking at the presence of URLs to fact-checking websites. Our findings il-

lustrated how crowds play the most prominent role in countering online misinformation—96% of all counter-misinformation messages are made by crowds. However, results showed that 2 out of 3 crowd-generated counter-response messages are rude and non-evidenced. Uncivil counter-responses can lead to reduced trust in the correcting user[67] and result in arguments.[68] We also found that counter-misinformation behavior is exhibited on multiple web and social media platforms, but is conducted inefficiently.[69] This implies a need to empower crowds so they counter misinformation more effectively.[70]

Other works have also conducted in-the-wild social media analysis of crowd fact-checking online. Vo et al.[71] identified fact-checking replies by checking whether a reply to misinformation contains a fact-checking URL towards one of two trustworthy fact-checking websites (i.e., Snopes.com and Politifact.com). Then, they retrieved the corresponding misinformation tweet toward which the fact-checking post replies, and use them to construct pairs of misinformation posts and fact-checking replies for fact-checking content analysis and reply generation. Miyazaki et al.[72] curated a large-scale dataset containing pairs of misinformation tweets and debunking replies, by first crawling COVID-19 related misinformation tweets from existing research[73] and then recruiting crowd-sourcing workers via Amazon Mechanical Turk to annotate responses to these tweets as being debunking or not. They finally performed analysis to illustrate who counters misinformation and how they do so.

Together, these works show that crowds actively engage in fact-checking and social correction of misinformation. This is a promising sign. We need more investment and research into making social correction effective and actionable for everyone. That would be an important step towards creating a self-correcting society that can be resilient against misinformation.

Which Misinformation do Crowds Correct? We recently investigated the properties of misinformation posts that get corrected.[74] Given the lack of a systematic misinformation correction by crowds, the characteristics of tweets that attract social correction from crowds versus those that do not remain unknown. Our work answered the following two research questions: (1) "Given a misinformation tweet, will it be countered by other users?" and (2) "If yes, what will be the magnitude of countering it?" This exploration can help develop instruments on guiding users' misinformation correction efforts and to measure disparity across users who get corrected. We created a novel Twitter dataset consisting of 690,047 pairs of misinformation tweets and counter-misinformation replies. Next, the stratified analysis of tweet linguistic and engagement features as well as tweet posters' user attributes was performed to demonstrate the factors that are influential in deciding whether or not a tweet will get countered. We found that misinformation tweets expressing negative sentiment, strong emotion, third-person pronouns, and impolite strategies are more likely to result in more countering replies from users. Predictive classifiers were finally created to predict the likelihood of a misinformation tweet to get countered and the degree to which that tweet will be countered.[75]

Empowering Crowds to Effectively Counter Misinformation with Generative AI Since 2 out 3 crowd correctors are rude and non-evidenced,[76] it is essential to improve the quality of crowd-generated misinformation corrections for the corrections to be effective. Uncivil corrections can backfire. Towards this goal, with our collaborators, we created an AI model called MisinfoCorrect to generate factual counter-responses to misinformation posts[77] in order to assist ordinary users reply with factual, polite, and countering

responses. We created two novel datasets of misinformation and counter-misinformation response pairs from in-the-wild social media and crowdsourcing from college-educated students. Using annotations on the collected data, poor responses were distinguished from ideal responses that are factual, polite, and refute misinformation. The work proposes MisinfoCorrect, a reinforcement learning-based framework that learns to generate polite, factual, and refuting counter-misinformation responses for a misinformation post. Quantitative and qualitative evaluation showed that our model outperforms several baselines by generating high-quality counter-responses.[78]

Developing tools to counter online misinformation Our research is also developing misinformation mapping tools for experts to track the spread of misinformation and create interventions to prevent its spread further. The goal of these tools is to map the spreaders and consumers of misinformation, provide an easy-to-use interface to professionals to track and visualize the spread in real-time, and then enable them to deliver interventions to the users who are most vulnerable to misinformation. These tools will enable faster, accurate, and focused fact-checking by experts, as well as correction of misinformation.

Conclusion

Crowds are a promising and scalable solution to detect and correct misinformation. Current research has provided compelling evidence regarding the prevalence and effectiveness of crowds in flagging misinformation and providing corrections. These have been shown in in-lab experiments as well as in-the-wild social media settings.

Gaps remain in bringing the best practices of flagging and corrections to the crowds, enabling crowds to do it effectively, building tools that can help them do so, and engaging users to be proactive about

participating in these activities. A Wikipedia-like model which is governed, maintained, and run by crowds is a desirable scenario to improve the online information ecosystem and improve the quality of information integrity.

The current state of crowd-powered misinformation detection and flagging is promising. However, social correction needs to be made effective and actionable for everyone. More investment and research is needed to find the correct ways to make that happen. Crowd-powered solutions are an important step towards creating a self-correcting society that can be resilient against misinformation. The recent advances in generative AI and Large Language Models show importance in different fields including question answering and dialog systems. Using generative AI and large language models also provides a promising direction in combatting misinformation. These new technologies can be used to democratize misinformation correction. By arming crowds with the effective counter-responses that amplify fact checks has the potential to dramatically reduce the uptake of misinformation, even in networks that are traditionally siloed from fact checks. The power and potential to use generative AI for social good is ripe for disruption.

Acknowledgements

This research/material is based upon work supported in part by NSF grants CNS-2154118, IIS-2027689, ITE-2137724, ITE-2230692, CNS-2239879 (NSF CAREER), Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112290102 (subcontract No. PO70745), and funding from Microsoft, Google, and Adobe Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily

reflect the position or policy of DARPA, DoD, NSF, and SRI International and no official endorsement should be inferred.

Notes

1 Francesco Pierri, *Online Misinformation Is Linked to Early COVID-19 Vaccination Hesitancy and Refusal*, 12 Sci. Reps. 1 (2022); Philip Ball & Amy Maxmen, *The Epic Battle Against Coronavirus Misinformation and Conspiracy Theories* (May 2020); David M.J. Lazer et al., *The Science of Fake News*, 359 Science 1094 (2018); Srijan Kumar, *Advances in AI for Web Integrity, Equity, and Well-Being*, 6 Frontiers Big Data 33 (2023).

2 Kate Starbird, *Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter After the 2013 Boston Marathon Bombing*, *in* 2014 IConference Proc.; Ahmer Arif, Leo Graiden Stewart & Kate Starbird, *Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse*, *in* 2 Proc. ACM on Hum.-Comput. Interaction (Nov. 2018), https://doi.org/10.1145/3274289.

3 Craig Silverman, *Lies, Damn Lies, and Viral Content: How News Websites Spread (and Debunk) Online Rumors, Unverified Claims and Misinformation*, 168 Tow Ctr. for Digit. Journalism 134 (2015) [hereinafter Silverman, *Lies*]; Craig Silverman, *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook*, BuzzFeed News (2016) [hereinafter Silverman, *Analysis*], https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook; Jieun Shin & Kjerstin Thorson, *Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media*, 67 J. Commc'n 233 (Feb. 2017), https://doi.org/10.1111/jcom.12284.

4 Leo G. Stewart, Ahmer Arif & Kate Starbird, *Examining Trolls and Polarization with a Retweet Network*, *in* 70 Proc. ACM Workshop on Misinformation & Misbehavior Mining on Web (2018).

5 Gaurav Verma, *Examining the Impact of Sharing COVID-19 Misinformation Online on Mental Health*, 12 Sci. Reps. 1 (2022).

6 Mason Walker & Katerina Eva Matsa, *News Consumption Across Social Media in 2021* (2021).

7 Soroush Vosoughi, Deb Roy & Sinan Aral, *The Spread of True and False News Online*, 359 Science 1146 (2018); Lazer et al., *supra* note 1.

8 Christian Fuchs, *Communicating COVID-19: Everyday Life, Digital Capitalism, and Conspiracy Theories in Pandemic Times* (2021).

9 Silverman, *Lies*, *supra* note 3; Silverman, *Analysis*, *supra* note 3; Shin & Thorson, *supra* note 3.

10 Starbird, *supra* note 2; Arif, Stewart & Starbird, *supra* note 2.

11 Stephan Lewandowsky, *Misinformation and Its Correction: Continued Influence and Successful Debiasing*, 13 Psych. Sci. Pub. Int. 106 (2012); Mahak Goindani & Jennifer Neville, *Social Reinforcement Learning to Combat Fake News Spread*, *in* 115 Proc. Uncertainty A.I. Conf. 1006 (Jan. 2020), https://proceedings.mlr.press/v115/goindani20a.html; Ceren Budak, Divyakant Agrawal & Amr El Abbadi, *Limiting the Spread of Misinformation in Social Networks*, *in* 20 Proc. Int'l Conf. on World Wide Web 665 (2011), http://portal.acm.org/citation.cfm?doid=1963405.1963499; Jianming Zhu, Smita Ghosh & Weili Wu, *Robust Rumor Blocking Problem with Uncertain Rumor Sources in Social Networks*, 24 World Wide Web 229 (Jan. 2021), https://link.springer.com/10.1007/s11280-020-00841-8; Iouliana Litou, *Efficient and Timely Misinformation Blocking Under Varying Cost Constraints*, 2 Online Soc. Networks & Media 19 (Aug. 2017), https://linkinghub.elsevier.com/retrieve/pii/S2468696417300113; Zhihong Wang & Yi Guo, *Rumor Events Detection Enhanced by Encoding Sentimental Information into Time Series Division and Word Representations*, 397 Neurocomputing 224 (July 2020), https://linkinghub.elsevier.com/retrieve/pii/S0925231220301533.

12 Srijan Kumar & Neil Shah, False Information on Web and Social Media: A Survey (2018) (unpublished manuscript), https://arxiv.org/abs1804.08559; Liang Wu, *Misinformation in Social Media: Definition, Manipulation, and Detection*, 21 ACM SIGKDD Expls. Newsl. 80 (2019).

13 Md Rafiqul Islam, *Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives*, 10 Soc. Network

Analysis & Mining 1 (2020); Jevin D. West & Carl T. Bergstrom, *Misinformation in and About Science*, *in* 118 Proc. Nat'l Acad. Scis. (2021); Xinyi Zhou & Reza Zafarani, *A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities*, 53 ACM Comput. Surveys 1 (2020); Kai Shu, *Fake News Detection on Social Media: A Data Mining Perspective*, 19 ACM SIGKDD Expls. Newsl. 22 (2017) [hereinafter Shu, *Fake News*]; Lazer et al., *supra* note 1.

14 Shu, *Fake News*, *supra* note 13; Kumar & Shah, *supra* note 12; Ray Oshikawa, Jing Qian & William Yang Wang, *A Survey on Natural Language Processing for Fake News Detection*, *in* 12 Proc. Language Res. & Evaluation Conf. 6086 (2020); Islam, *supra* note 13; Zhou & Zafarani, *supra* note 13.

15 Aditi Gupta, *Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy*, *in* 22 Proc. Int'l Conf. on World Wide Web 729–736 (2013), https://doi.org/10.1145/2487788.2488033; Benjamin D. Horne & Sibel Adali, *This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News*, *in* 11 Int'l AAAI Conf. on Web & Soc. Media (2017); Philip N. Howard & Bence Kollanyi, Bots,# StrongerIn, and# Brexit: Computational Propaganda During the UK-EU Referendum (2016) (unpublished manuscript), https://arxiv.org/abs1606.06356; Vahed Qazvinian, *Rumor Has It: Identifying Misinformation in Microblogs*, *in* 2011 Proc. Conf. on Empirical Methods Nat. Language Processing 1589; Tanushree Mitra & Eric Gilbert, *Credbank: A Large-Scale Social Media Corpus with Associated Credibility Annotations*, *in* 9 Int'l AAAI Conf. on Web & Soc. Media (2015); Tanushree Mitra, Graham P. Wright & Eric Gilbert, *A Parsimonious Language Model of Social Media Credibility Across Disparate Events*, *in* 2017 Proc. ACM Conf. on Comput. Supported Coop. Work & Soc. Comput. 126; Verónica Pérez-Rosas, *Automatic Detection of Fake News*, *in* 27 Proc. Int'l Conf. on Comput. Linguistics 3391 (2018).

16 Eugenio Tacchini, Some like It Hoax: Automated Fake News Detection in Social Networks (2017) (unpublished manuscript), https://arxiv.org/abs1704.07506; Gupta, *supra* note 15; Clayton Allen Davis, *Botornot: A System to Evaluate Social Bots*, *in* 25 Proc. Int'l Conf. Companion on World

Wide Web 273 (2016); Vosoughi, Roy & Aral, *supra* note 7; Chengcheng Shao, *Hoaxy: A Platform for Tracking Online Misinformation*, *in* 25 Proc. Int'l Conf. Companion on World Wide Web 745–750 (2016), https://doi.org/10.1145/2872518.2890098.

17 Tacchini, *supra* note 16; Alessandro Bessi & Emilio Ferrara, *Social Bots Distort the 2016 US Presidential Election Online Discussion*, *in* 21 First Monday (2016); Davis, *supra* note 16; Adrien Friggeri, *Rumor Cascades*, *in* 8 Proc. Int'l AAAI Conf. on Web & Soc. Media 101 (2014); Marcelo Mendoza, Barbara Poblete & Carlos Castillo, *Twitter Under Crisis: Can We Trust What We RT?*, *in* 1 Proc. Workshop on Soc. Media Analytics 71 (2010); Qazvinian, *supra* note 15; A. Conrad Nied, *Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media*, *in* 2017 Companion ACM Conf. on Comput. Supported Coop. Work & Soc. Comput. 263; Kate Starbird, *Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter*, *in* 11 Int'l AAAI Conf. on Web & Soc. Media (2017); Venkatramanan S. Subrahmanian, *The DARPA Twitter Bot Challenge*, 49 Computer 38 (2016).

18 Shao, *supra* note 16; Friggeri, *supra* note 17; Vosoughi, Roy & Aral, *supra* note 7; Savvas Zannettou, *The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources*, *in* 2017 Proc. Internet Measurement Conf. 405.

19 Kim L. Fridkin, *Gender Differences in Reactions to Fact Checking of Negative Commercials*, 12 Pol. & Gender 369 (June 2016); Fang Jin, *Epidemiological Modeling of News and Rumors on Twitter*, *in* 7 Proc. Workshop on Soc. Network Mining & Analysis 1 (2013); Shao, *supra* note 16; Chengcheng Shao, *The Spread of Low-Credibility Content by Social Bots*, 9 Nature Commc'ns 1 (2018); Arkaitz Zubiaga, *Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads*, *in* 11 PloS One (2016); Vosoughi, Roy & Aral, *supra* note 7.

20 https://www.snopes.com/

21 John Zarocostas, *How to Fight an Infodemic*, 395 Lancet 676 (2020).

22 Jennifer Allen, *Scaling up Fact-Checking Using the Wisdom of Crowds*, *in* 7 Sci. Advances No. 36 (Sept. 2021), https://www.science.org/doi/10.1126/sciadv.abf4393; Nicholas Micallef, *The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic*, *in* 2020 IEEE Int'l Conf. on Big Data [hereinafter Micallef, *Role*]; Jooyeon Kim, *Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation*, *in* 11 Proc. ACM Int'l Conf. on Web Search & Data Mining 324 (2018); Kevin Roitero, *Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background*, *in* 43 Proc. Int'l ACM SIGIR Conf. on Rsch. & Dev. Info. Retrieval 439 (2020); Haeseung Seo, *If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation*, *in* 16 Proc. Int'l AAAI Conf. on Web & Soc. Media 896 (2022) [hereinafter Seo, *Reliable Source*].

23 Bing He, Mustaque Ahamad & Srijan Kumar, *Petgen: Personalized Text Generation Attack on Deep Sequence Embedding-Based Classification Models*, *in* 27 Proc. ACM Conf. on Knowledge Discovery & Data Mining 575 (2021).

24 Gautam Kishore Shahi, Anne Dirkson & Tim A. Majchrzak, *An Exploratory Study of COVID-19 Misinformation on Twitter*, 22 Online Soc. Networks & Media 100104 (2021); Allen, *supra* note 22.

25 Melissa Tully, Leticia Bode & Emily K. Vraga, *Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users' Responses to a Health Misinformation Post?*, 6 Soc. Media + Soc'y 205630512097837 (Oct. 2020), http://journals.sagepub.com/doi/10.1177/2056305120978377; Nguyen Vo & Kyumin Lee, *The Rise of Guardians: Fact-Checking Url Recommendation to Combat Fake News*, *in* 41 Int'l ACM Conf. on Rsch. & Dev. Info. Retrieval 275 (2018); Xinyi Zhou, *"This Is Fake! Shared It by Mistake": Assessing the Intent of Fake News Spreaders*, *in* 2022 Proc. ACM Web Conf. 3685; Leticia Bode & Emily K. Vraga, *See Something, Say Something: Correction of Global Health Misinformation on Social Media*, 33 Health Commc'n 1131 (Sept. 2018) [hereinafter Bode & Vraga, *See Something*], https://www.tandfonline.com/doi/full/10.1080/10410236.2017.1331312; Starbird, *supra* note 2; Kunihiro Miyazaki, "This Is Fake News": Characterizing the Spontaneous De-

bunking from Twitter Users to COVID-19 False Information (2022) (unpublished manuscript), https://arxiv.org/abs2203.14242; Yingchen Ma, *Characterizing and Predicting Social Correction on Twitter*, *in* 15 ACM Web Sci. Conf. (2023).

26 Allen, *supra* note 22; Kim, *supra* note 22.

27 Van-Hoang Nguyen, *Fang: Leveraging Social Context for Fake News Detection Using Graph Representation*, *in* 29 Proc. ACM Int'l Conf. on Info. & Knowledge Mgmt. 1165 (2020); Kai Shu, Suhang Wang & Huan Liu, *Beyond News Contents: The Role of Social Context for Fake News Detection*, *in* 12 Proc. ACM Int'l Conf. on Web Search & Data Mining 312 (2019); Zhiwei Jin, *News Credibility Evaluation on Microblog with a Hierarchical Propagation Model*, *in* 2014 IEEE Int'l Conf. on Data Mining 230; Kai Shu, *Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation*, *in* 14 Proc. Int'l AAAI Conf. on Web & Soc. Media 626 (2020) [hereinafter Shu, *Hierarchical*].

28 Micallef, *Role*, *supra* note 22.

29 Seo, *Reliable Source*, *supra* note 22.

30 Friggeri, *supra* note 17; Jonas Colliander, *"This Is Fake News": Investigating the Role of Conformity to Other Users' Views when Commenting on and Spreading Disinformation in Social Media*, 97 Computs. Hum. Behav. 202 (2019); Senuri Wijenayake, *Effect of Conformity on Perceived Trustworthiness of News in Social Media*, 25 IEEE Internet Comput. 12 (Jan. 2021), https://ieeexplore.ieee.org/document/9233938/.

31 Leticia Bode, Emily K. Vraga & Melissa Tully, *Do the Right Thing: Tone May Not Affect Correction of Misinformation on Social Media*, *in* 2020 Harv. Kennedy Sch. Misinformation Rev.; Emily K. Vraga & Leticia Bode, *I Do Not Believe You: How Providing a Source Corrects Health Misperceptions Across Social Media Platforms*, 21 Info. Commc'n & Soc'y 1337 (2018) [hereinafter Vraga & Bode, *I Do Not Believe You*]; Emily K. Vraga & Leticia Bode, *Addressing COVID-19 Misinformation on Social Media Preemptively and Responsively*, 27 Emerging Infectious Diseases 396 (2021) [hereinafter Vraga & Bode, *Addressing*]; Leticia Bode & Emily K. Vraga, *In Related News,*

That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media, 65 J. Commc'n 619 (June 2015) [hereinafter Bode & Vraga, *In Related News*], https://doi.org/10.1111/jcom.12166; Bode & Vraga, *See Something*, *supra* note 25; Emily K. Vraga & Leticia Bode, *Correction as a Solution for Health Misinformation on Social Media*, 110 Am. J. Pub. Health S278 (2020) [hereinafter Vraga & Bode, *Correction*].

32 Vraga & Bode, *Addressing*, *supra* note 31; Emily Vraga, Melissa Tully & Leticia Bode, *Assessing the Relative Merits of News Literacy and Corrections in Responding to Misinformation on Twitter*, 24 New Media & Soc'y 2354 (2021); Emily K. Vraga, *Testing the Effectiveness of Correction Placement and Type on Instagram*, 25 Int'l J. Press/Pol. 632 (2020); Emily K. Vraga, Leticia Bode & Melissa Tully, *The Effects of a News Literacy Video and Real-Time Corrections to Video Misinformation Related to Sunscreen and Skin Cancer*, 2021 Health Commc'n 1.

33 Seo, *Reliable Source*, *supra* note 22; Michael Hameleers & Toni G.L.A. van der Meer, *Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers?*, 47 Commc'n Rsch. 227 (2020); Michael Hameleers, *Separating Truth from Lies: Comparing the Effects of News Media Literacy Interventions and Fact-Checkers in Response to Political Misinformation in the US and Netherlands*, 25 Info. Commc'n & Soc'y 110 (2022); Jingwen Zhang, *Effects of Fact-Checking Social Media Vaccine Misinformation on Attitudes Toward Vaccines*, 145 Preventive Med. 106408 (2021); Nathan Walter, *Fact-Checking: A Meta-Analysis of What Works and for Whom*, 37 Pol. Commc'n 350 (2020); Tanja Pavleska, *Performance Analysis of Fact-Checking Organizations and Initiatives in Europe: A Critical Overview of Online Platforms Fighting Fake News*, 29 Soc. Media & Convergence 1 (2018).

34 Md Momen Bhuiyan, *Investigating Differences in Crowdsourced News Credibility Assessment*, 4 Proc. ACM on Hum.-Comput. Interaction 1 (Oct. 2020), https://dl.acm.org/doi/10.1145/3415164.

35 James Surowiecki, *The Wisdom of Crowds* (2005).

36 Allen, *supra* note 22; Michael Soprano, *The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale*, 58 Info. Processing & Mgmt. 102710 (Nov. 2021), https://linkinghub.elsevier.com/retrieve/pii/S0306457321001941.

37 Richard McCreadie, Craig Macdonald & Iadh Ounis, *Crowdsourced Rumour Identification During Emergencies*, *in* 24 Proc. Int'l Conf. on World Wide Web 965 (May 2015), https://dl.acm.org/doi/10.1145/2740908.2742573.

38 Allen, *supra* note 22; Soprano, *supra* note 36.

39 Gordon Pennycook & David G. Rand, *Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality*, 116 Proc. Nat'l Acad. Scis. 2521 (Feb. 2019), http://www.pnas.org/lookup/doi/10.1073/pnas.1806781116.

40 Shan Jiang, *Modeling and Measuring Expressed (Dis) Belief in (Mis) Information*, *in* 14 Proc. Int'l AAAI Conf. on Web & Soc. Media 315 (2020); Kai Shu, *Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media*, 8 Big Data 171 (2020); Micallef, *Role*, *supra* note 22; Nicholas Micallef, *Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos*, *in* 16 Proc. Int'l AAAI Conf. on Web & Soc. Media 651 (2022) [hereinafter Micallef, *Cross-Platform*]; Kai Shu, Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News (2020) [hereinafter Shu, *Leveraging*] (unpublished manuscript), https://arxiv.org/abs2004.01732; Nima Noorshams, Saurabh Verma & Aude Hofleitner, *TIES: Temporal Interaction Embeddings For Enhancing Social Media Integrity At Facebook*, *in* 26 Proc. ACM Conf. on Knowledge Discovery & Data Mining 3128 (2020); Fatema Hasan, Learning User Embeddings from Temporal Social Media Data: A Survey (2021) (unpublished manuscript), https://arxiv.org/abs/2105.07996; Nguyen, *supra* note 27; Shu, Wang & Liu, *supra* note 27; Jin, *supra* note 27; Shu, *Hierarchical*, *supra* note 27; Yingtong Dou, *User Preference-Aware Fake News Detection*, *in* 44 Proc. Int'l ACM SIGIR Conf. on Rsch. & Dev. Info. Retrieval 2051 (2021); Qazvinian, *supra* note 15; Shaina Raza & Chen Ding, *Fake News Detection Based on News Content*

and Social Contexts: A Transformer-Based Approach, 13 Int'l J. Data Sci. & Analytics 335 (2022); Natali Ruchansky, Sungyong Seo & Yan Liu, *Csi: A Hybrid Deep Model for Fake News Detection*, *in* 2017 Proc. ACM on Conf. on Info. & Knowledge Mgmt. 797.

41 Nguyen, *supra* note 27; Shu, Wang & Liu, *supra* note 27; Jin, *supra* note 27; Shu, *Hierarchical*, *supra* note 27; Dou, *supra* note 40; Shu, *Leveraging*, *supra* note 40; Qazvinian, *supra* note 15; Raza & Ding, *supra* note 40; Ruchansky, Seo & Liu, *supra* note 40; Noorshams, Verma & Hofleitner, *supra* note 40.

42 Shu, *Fake News*, *supra* note 13; Kumar & Shah, *supra* note 12; Oshikawa, Qian & Wang, *supra* note 14; Islam, *supra* note 13; Zhou & Zafarani, *supra* note 13.

43 Jennifer Allen, Cameron Martel & David G. Rand, *Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program*, *in* 2022 CHI Conf. on Hum. Factors Comput. Sys. 1; Rohit Mujumdar & Srijan Kumar, *HawkEye: A Robust Reputation System for Community-Based Counter-Misinformation*, *in* 2021 Proc. IEEE/ACM Int'l Conf. on Advances Soc. Networks Analysis & Mining 188.

44 Allen, Martel & Rand, *supra* note 43.

45 Nicolas Pröllochs, *Community-Based Fact-Checking on Twitter's Birdwatch Platform*, *in* 16 Proc. Int'l AAAI Conf. on Web & Soc. Media 794 (2022); Miyazaki, *supra* note 25; Chiara Drolsbach & Nicolas Pröllochs, Diffusion of Community Fact-Checked Misinformation on Twitter (2022) (unpublished manuscript), https://arxiv.org/abs/20205.13673; Allen, Martel & Rand, *supra* note 43.

46 Mujumdar & Kumar, *supra* note 43.

47 Mujumdar & Kumar, *supra* note 43.

48 Porismita Borah, Bimbisar Irom & Ying Chia Hsu, *"It Infuriates Me": Examining Young Adults' Reactions to and Recommendations to Fight Misinformation About COVID-19*, Aug. 2021 J. Youth Stud. 1, https://www.

tandfonline.com/doi/full/10.1080/13676261.2021.1965108; Jan Kirchner & Christian Reuter, *Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness*, 4 Proc. ACM on Hum.-Comput. Interaction 1 (Oct. 2020), https://dl.acm.org/doi/10.1145/3415211; Tully, Bode & Vraga, *supra* note 25.

49 Jeyasushma Veeriah, *Young Adults' Ability to Detect Fake News and Their New Media Literacy Level in the Wake of the Covid-19 Pandemic*, 13 J. Content Cmty. & Commc'n 372 (2021); Kirchner & Reuter, *supra* note 48.

50 Tully, Bode & Vraga, *supra* note 25.

51 Seo, *Reliable Source*, *supra* note 22.

52 Friggeri, *supra* note 17; Colliander, *supra* note 30; Wijenayake, *supra* note 30.

53 Bode, Vraga & Tully, *supra* note 31; Vraga & Bode, *I Do Not Believe You*, *supra* note 31; Vraga & Bode, *Addressing*, *supra* note 31; Bode & Vraga, *In Related News*, *supra* note 31; Bode & Vraga, *See Something*, *supra* note 25; Vraga & Bode, *Correction*, *supra* note 31.

54 Vraga & Bode, *Addressing*, *supra* note 31; Vraga, Tully & Bode, *supra* note 32; Vraga, *supra* note 32; Vraga, Bode & Tully, *supra* note 32.

55 Maryke S. Steffens, *How Organisations Promoting Vaccination Respond to Misinformation on Social Media: A Qualitative Investigation*, 19 BMC Pub. Health 1 (2019); Pranav Malhotra, Kristina Scharp & Lindsey Thomas, *The Meaning of Misinformation and Those Who Correct It: An Extension of Relational Dialectics Theory*, 39 J. Soc. & Pers. Relationships 1256 (2022); Yuko Tanaka & Rumi Hirayama, *Exposure to Countering Messages Online: Alleviating or Strengthening False Belief?*, 22 Cyberpsychology Behav. & Soc. Networking 742 (Nov. 2019), https://www.liebertpub.com/doi/10.1089/cyber.2019.0227; Seo, *Reliable Source*, *supra* note 22; Gábor Orosz, *Changing Conspiracy Beliefs Through Rationality and Ridiculing*, *in* 7 Frontiers Psych. (Oct. 2016); Ana Stojanov, *Reducing Conspiracy Theory Beliefs*, 48 Psihologija 251 (2015); Man-pui Sally Chan, *Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation*, 28 Psych. Sci. 1531 (2017); Bing He, Mustaque Ahamad &

Srijan Kumar, *Reinforcement Learning-Based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation*, *in* 2023 Proc. ACM Web Conf..

56 Chan, *supra* note 55; Nathan Walter, *Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis*, 36 Health Commc'n 1776 (2021); Walter, *supra* note 33; Nathan Walter & Sheila T. Murphy, *How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation*, 85 Commc'n Monographs 423 (2018); Ethan Porter & Thomas J. Wood, *The Global Effectiveness of Fact-Checking: Evidence from Simultaneous Experiments in Argentina, Nigeria, South Africa, and the United Kingdom*, *in* 118 Proc. Nat'l Acad. Scis. e2104235118 (2021).

57 Briony Swire-Thompson, Joseph DeGutis & David Lazer, *Searching for the Backfire Effect: Measurement and Design Considerations*, 9 J. Applied Rsch. Memory & Cognition 286 (2020); Andrew Guess & Alexander Coppock, *Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments*, 50 British J. Pol. Sci. 1497 (2020); Thomas Wood & Ethan Porter, *The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence*, 41 Pol. Behav. 135 (2019).

58 Drew B. Margolin, Aniko Hannak & Ingmar Weber, *Political Fact-Checking on Twitter: When Do Corrections Have an Effect?*, 35 Pol. Commc'n 196 (2018).

59 Friggeri, *supra* note 17; Veeriah, *supra* note 49; Sukeshini Grandhi, Linda Plotnick & Starr Roxanne Hiltz, *By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media*, 5 Proc. ACM on Hum.-Comput. Interaction 1 (July 2021), https://dl.acm.org/doi/10.1145/3463930; Mohsen Mosleh, *Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment*, *in* 2021 Proc. CHI Conf. on Hum. Factors Comput. Sys. 1; Anjan Pal et al., *Rumor Analysis & Visualization System*, *in* 2019 Proc. Int'l Multi Conf. Eng'rs & Comput. Scientists.

60 Yanqing Sun, *The Role of Influence of Presumed Influence and Anticipated Guilt in Evoking Social Correction of COVID-19 Misinformation*, Feb. 2021 Health Commc'n 1, https://www.tandfonline.com/doi/full/10.1080/10410236.2021.1888452.

61 Leticia Bode & Emily K. Vraga, *Correction Experiences on Social Media During COVID-19*, 7 Soc. Media + Soc'y 205630512110088 (Apr. 2021) [hereinafter Bode & Vraga, *Correction Experiences*], http://journals.sagepub.com/doi/10.1177/20563051211008829; Colliander, *supra* note 30; Bode & Vraga, *See Something*, *supra* note 25; Wijenayake, *supra* note 30; Haeseung Seo, *(In) Effectiveness of Accumulated Correction on COVID-19 Misinformation*, *in* 2021 Tech. Mind & Soc'y Conf. Proc. [hereinafter Seo, *Effectiveness*].

62 He, Ahamad & Kumar, *supra* note 55.

63 Bode & Vraga, *Correction Experiences*, *supra* note 61; Colliander, *supra* note 30; Bode & Vraga, *See Something*, *supra* note 25; Wijenayake, *supra* note 30; Seo, *Effectiveness*, *supra* note 61.

64 Pröllochs, *supra* note 45.

65 Micallef, *Role*, *supra* note 22; Micallef, *Cross-Platform*, *supra* note 40.

66 Micallef, *Role*, *supra* note 22.

67 Lucie Flekova, Daniel Preoțiuc-Pietro & Lyle Ungar, *Exploring Stylistic Variation with Age and Income on Twitter*, *in* 54 Proc. Ann. Meeting Ass'n for Comput. Linguistics (Volume 2: Short Papers) 313 (2016); Kjerstin Thorson, Emily Vraga & Brian Ekdale, *Credibility in Context: How Uncivil Online Commentary Affects News Credibility*, 13 Mass Commc'n & Soc'y 289 (2010).

68 Gina M. Masullo & Jiwon Kim, *Exploring "Angry" and "Like" Reactions on Uncivil Facebook Comments That Correct Misinformation in the News*, 9 Digit. Journalism 1103 (Sept. 2021), https://www.tandfonline.com/doi/full/10.1080/21670811.2020.1835512; Srijan Kumar, *Community Interaction and Conflict on the Web*, *in* 2018 Proc. World Wide Web Conf. 933; Justin Cheng, *Anyone Can Become a Troll: Causes of Trolling Behavior in*

Online Discussions, *in* 2017 Proc. ACM Conf. on Comput. Supported Coop. Work & Soc. Comput. 1217.

69 Micallef, *Cross-Platform*, *supra* note 40.

70 The code and data for this work are present at Bing He, *Countering COVID-19 Misinformation Tweet Code & Data* (last updated Feb. 10, 2023), http://claws.cc.gatech.edu/covid_counter_misinformation/.

71 Nguyen Vo & Kyumin Lee, *Learning from Fact-Checkers: Analysis and Generation of Fact-Checking Language*, *in* 42 Proc. Int'l ACM SIGIR Conf. on Rsch. & Dev. Info. Retrieval 335 (2019).

72 Miyazaki, *supra* note 25.

73 Limeng Cui & Dongwon Lee, Coaid: Covid-19 Healthcare Misinformation Dataset (2020) (unpublished manuscript), https://arxiv.org/abs/2006.00885; Jisu Kim, *FibVID: Comprehensive Fake News Diffusion Dataset During the COVID-19 Period*, 64 Telematics & Informatics 101688 (2021); Tamanna Hossain, *COVIDLies: Detecting COVID-19 Misinformation on Social Media*, *in* 1 Proc. Workshop on NLP for COVID-19 (2020); Shahan Ali Memon & Kathleen M. Carley, Characterizing Covid-19 Misinformation Communities Using a Novel Twitter Dataset (2020) (unpublished manuscript), https://arxiv.org/abs2008.00791; Shahi, Dirkson & Majchrzak, *supra* note 24.

74 Ma, *supra* note 25.

75 The code and data for this work are present at Bing He & Srijan Kumar, *Characterizing and Predicting Social Correction on Twitter* (last updated May 20, 2023), https://github.com/claws-lab/social-correction-twitter.

76 Micallef, *Role*, *supra* note 22.

77 He, Ahamad & Kumar, *supra* note 55.

78 The code and data for this work are present at Bing He & Srijan Kumar, *Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation* (June 28, 2023), https://github.com/claws-lab/MisinfoCorrect.

Chapter 8

I'm Not a Robot: Using Authenticity to Govern User Behavior

Alice E. Marwick

Since the earliest precursors of today's social media platforms, people have classified online behaviors as desirable or undesirable. Community-written Frequently Asked Questions, or FAQs, provided answers to common questions to prevent people clogging up Usenet fora with repetitive inquiries. Administrators of early text-based games and chat spaces quarantined or removed users whose behavior offended others. In the 1990s, computer-mediated communication scholars investigated "flaming," trying to understand why arguments over email escalated rapidly to hateful and aggressive language. Over time, the implicit norms that existed in emergent social internet spaces codified into explicit guidelines and regulations. The laissez-faire attitudes of Web 2.0 have been replaced by platform community standards that govern everything from acceptable language to advertis-

ing to classifying certain behaviors as harmful. These documents are frequently referred to by Trust and Safety personnel as the "constitution" or "laws" of platforms, implying that content or behavior that falls short of these regulations will be quickly sanctioned. In practice, this process is difficult and often involves rapid subjective judgments made by teams of underpaid and under-resourced content moderators. Determining whether these decisions are right or wrong is frequently far from obvious.

Virtually every platform has rules on what constitutes "authentic" and "inauthentic" behavior, with the former held up as an ideal while the latter portrayed as dangerous or damaging. TikTok, for example, opens its community guidelines with a paean to the authentic:

> At TikTok, we prioritize safety, diversity, inclusion, and authenticity. We encourage creators to celebrate what makes them unique and viewers to engage with what inspires them; we believe that a safe environment helps everyone do so openly. We prize the global nature of our community and strive to take into account the breadth of cultural norms where we operate. We also aim to cultivate an environment for genuine interactions by encouraging authentic content on TikTok.[1]

But authenticity is a slippery term. Alone, it doesn't mean much; it's always judged relative to something else that is *in*authentic.[2] Some music is authentic while other music is not; some people perform authentically online while others do not. Authenticity has become a primary organizing principle of online content in general. Highly valued content is authentic (truthful, expressive), and creators deemed authentic can reap the benefits of an engaged and active audience.[3] "Authentic" YouTubers, TikTokers, and Instagram stars are often contrasted to inauthentic "broadcast" or "legacy" celebrities. In this case, "inauthentic" celebrities are those who carefully craft their images, conceal

their "real selves" from the public, and are paid to promote products even if they do not like or use them, while "authentic" online celebrities are who they appear to be and only accept sponsorship from products they endorse (this is, of course, a false distinction). When people praise certain creators for their "authenticity," they are really praising a successful *performance* of authenticity. All content creators make careful decisions as to what they reveal and conceal to their followers or fans. Even content that looks spontaneous, messy, or potentially damaging can be carefully designed to elicit an affective response between creator and audience.

Authentic behavior, on the other hand, is something else. Social platforms define authenticity differently. Facebook writes, on its Community Standards page:

> In line with our commitment to authenticity, we do not allow people to misrepresent themselves on Facebook, use fake accounts, artificially boost the popularity of content or engage in behaviors designed to enable other violations under our Community Standards. This policy is intended to protect the security of user accounts and our services, and create a space where people can trust the people and communities they interact with.[4]

In enforcing its commitment to "authenticity," Facebook famously has a "real names" policy in which users cannot use names other than their legal name. This is automatically enforced by Facebook's moderation software, causing enormous problems for several distinct groups of users. These include trans people who may go by a different name than their legal name, people from cultures whose naming conventions differ from the Western firstname lastname patronymic standard, or people whose names are in languages like Scottish Gaelic, which contain capitalization or characters that differ from standard

English. Facebook does not allow people to have parody accounts or accounts for characters or celebrities unless they are run by official channels. Such accounts are common on sites like Twitter and were quaintly called "Fakesters" on the early 2000s social networking site Friendster.[5] (Friendster's ban of such accounts, along with site performance issues, led users to leave in droves for MySpace and eventually the site's demise.) Today's social platforms fall along a spectrum of performativity, playfulness, and anonymity, with Facebook and LinkedIn on one end, and Reddit, Twitter, and Instagram on the other. It's very common to have multiple accounts on Reddit and Twitter, and Instagram accounts can be anonymous (such as the gossip account Deux-Moi) or run by teams of individuals.

There are many advantages to persistent pseudonyms. People feel more comfortable sharing sensitive information in online support groups when it is not tied to their "real names," but find it easier to trust other posters if they recognize their usernames.[6] Pseudonyms prevent context collapse, as they allow people to create separate identities for their different internet pastimes: one user might use their "real name" for professional interactions on LinkedIn, a handle for fannish interactions on Discord or Archive of our Own, and a nickname for an Instagram account followed by their high school friends. This is not inauthentic behavior and there is nothing suspicious about it. While supporters of "real name" policies often claim that people who use their legal name are less likely to engage in antisocial or harassing behavior, there is no clear empirical evidence of this.[7] There is certainly plenty of incivility on Facebook. Facebook's hard stance conflating authenticity with legal names perhaps makes more sense in the context of online advertising, as maintaining multiple disconnected personae makes it more difficult for advertisers to create persistent profiles of users as they move around the web.

The other major regulation of authentic and inauthentic behavior by social platforms involves bots. Twitter's Spam Policy is a good example of this, as it includes prohibitions against "platform manipulation": "using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience." The company disallows "inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are" and "coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting."[8] In other words, bots that artificially inflate the number of followers an account has are problematic because they deceive the user into thinking an account is popular when it may not be; bots that engage in what Sam Woolley and Phillip Howard call "computational propaganda" are problematic because they may manipulate public opinion by making some points of view seem more popular than they actually are.[9]

This, of course, is the point of marketing. A handbag company who sends free purses to Instagram influencers is attempting to manipulate public opinion by making a purse seem fashionable and desired. Twitter or Facebook algorithmically promoting sponsored content is hoping that more people will see it than those who might "organically" seek it out. Political elites who spread negative information about their opponents are also "artificially influencing conversations." After all, the difference between propaganda and public relations is normative. (Americans generally see nothing wrong with the U.S. Air Force's involvement in 2022's *Top Gun: Maverick*, while China's equivalent, *Born to Fly*, is viewed suspiciously as blatant propaganda.)[10] In Twitter's case, the difference between *authentic* and *inauthentic* is not the behavior of the account, but the humanity of the creator. A bot is a piece of software, whereas an authentic user is a person. But it

can be very difficult to differentiate the two. Programs like Botometer and Bot Sentinel which attempt to automatically detect bots frequently fail, mislabeling humans as bots and vice versa. One study concluded that "it is impossible to say with absolute certainty that an account is a bot."[11] Drawing a clear, bright line between bot or not is thus bound to fail. Still, regulators are anxious to classify all bot activity as inauthentic while classifying similar efforts from paid advertisers or domestic governments as authentic, and thus allowable.

Using a slippery, normative concept like authenticity as an organizing principle to govern user behavior is misguided. Instead, platforms should focus on unwanted *behavior*, prioritizing the prevention of harm. Our concern over Russian Twitter accounts attempting to inflame racial divisions amongst Americans should not be greater than our concern over American Twitter users that spread "authentic" racist sentiments.[12] Authenticity sounds nice, but when it is leveraged by giant social platforms to justify contentious content moderation decisions, we should question the sentiments behind it.

Notes

1 *Community Guidelines*, TikTok (Oct. 2022), https://www.tiktok.com/community-guidelines.

2 *See* D. Grazian, *Blue Chicago: The Search for Authenticity in Urban Blues Clubs* (2003).

3 *See* Brooke Erin Duffy, *(Not) Getting Paid to Do What You Love: Gender, Social Media, and Aspirational Work* (2017); Sarah Banet-Weiser, *Authentic TM: The Politics of Ambivalence in a Brand Culture* (2012).

4 *See Inauthentic Behavior*, Facebook Cmty. Standards (2022), https://transparency.fb.com / policies / community - standards / inauthentic - behavior/.

5 *See* Alice Marwick, *"I'm a Lot More Interesting than a Friendster Profile": Identity Presentation, Authenticity and Power in Social Networking Services*, *in* 6 Ass'n Internet Researchers (2005).

6 *See* Joseph B. Walther & Shawn Boyd, *Attraction to Computer-Mediated Social Support*, *in* 2 Commc'n Tech. & Soc'y: Audience Adoption & Uses No. 153188 (2002).

7 *See* Kevin Munger, *Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment*, 39 Pol. Behav. 629 (2017), https://doi.org/10.1007/s11109-016-9373-5.

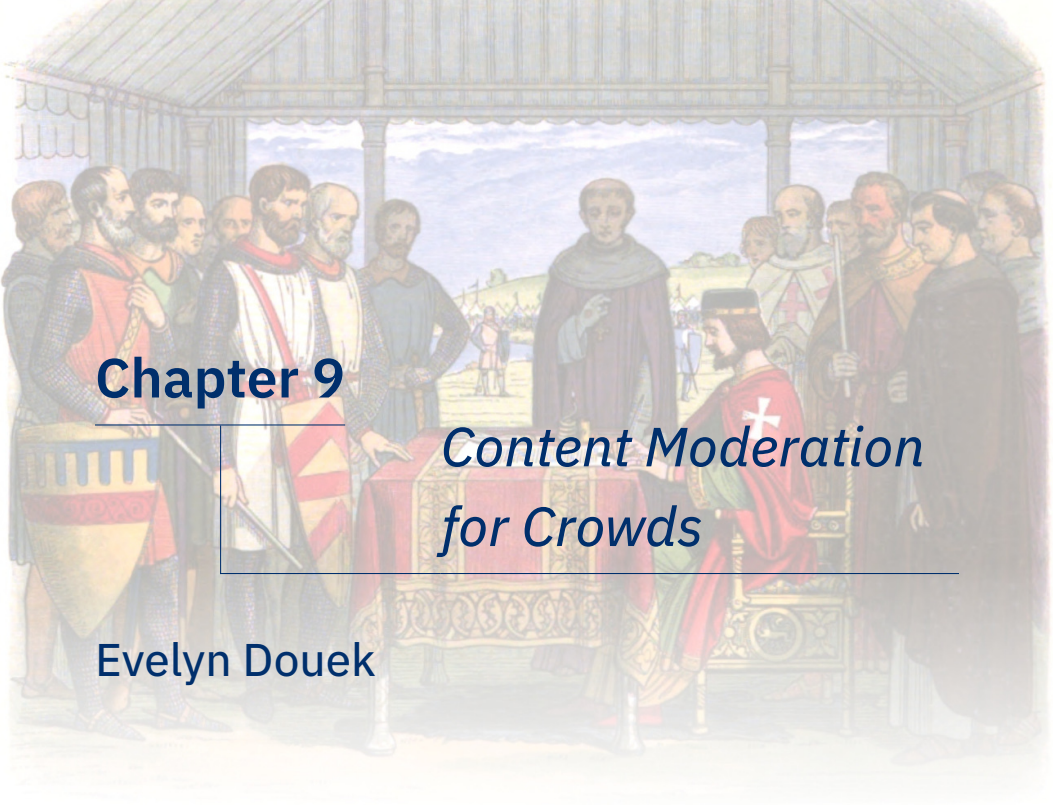
8 *Twitter's Platform Manipulation and Spam Policy*, Twitter Help Ctr. (Apr. 2022), https://help.twitter.com/en/rules-and-policies/platform-manipulation.

9 *See* Samuel C. Woolley & Philip N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media* (2018).

10 *Compare "Top Gun," Brought to You by the U.S. Military*, Wash. Post (May 27, 2022), https://www.washingtonpost.com/history/2022/05/27/top-gun-maverick-us-military/, *with* Vivienne Chow, *China Returns to Busan Film Festival*, Variety (blog) (Oct. 8, 2022), https://variety.com/2022/film/global/china-film-co-production-corp-busan-film-festival-1235398987/.

11 *See* Oliver Beatson et al., *Automation on Twitter: Measuring the Effectiveness of Approaches to Bot Detection*, Soc. Sci. Comput. Rev. (Aug. 6, 2021), https://doi.org/10.1177/08944393211034991.

12 *See* Deen Freelon et al., *Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation*, 40 Soc. Sci. Comput. Rev. 560 (2020), https://journals.sagepub.com/doi/abs/10.1177/0894439320914853.



Chapter 9

Content Moderation for Crowds

Evelyn Douek

In the past few years, platforms have increasingly been turning to performing content moderation on the basis of how people *act* and *associate* in groups, rather than on the basis of the *content* of particular posts. Two factors have led to this shift. First, as moderation based on the content of posts has become increasingly politically fraught, moderating on the basis of signals about a user's behavior or association appears to allow platforms to make decisions imposing community norms without having to pass judgment on particular speech. Second, platforms started to pay greater attention to harms caused by people acting in groups that are not visible or salient at the level of individual posts.

The first behavioral and associational moderation came out of the demand that platforms deal with information operations on their services after the 2016 U.S. election.[1] Platforms that had previously relied on community standards that focused on the content of particular posts found themselves without an existing rule addressing the way

Russian agents had taken advantage of their services. The harm of information operations happens in the aggregate: each individual post might not be particularly objectionable, but a campaign *as a whole* violates community norms. One inauthentic post by a Russian troll suggesting that Satan was rooting for Hillary Clinton in 2016 probably doesn't matter;[2] a slew of them might start to sow division or at the very least create a false impression of public sentiment. As Facebook's head of cybersecurity policy explained, "most of the content shared by coordinated manipulation campaigns isn't provably false, and would in fact be acceptable political discourse if shared by authentic audiences."[3] As a result they had to adopt a policy that was "based on the behavior we see on our platform—not based on who the actors behind it are or what they say."[4] Other platforms adopted similar approaches.[5] This had the advantage of allowing platforms to insist that they didn't take down information operations because of their political *content*, but because of their inauthentic behavior.

Increased crowd-based moderation may not only respond to a policy gap in platform rules, but also may be a practical necessity. Content moderation at scale cannot be done post-by-post. The volume of content moderation is hard to comprehend. The largest platforms take down tens of thousands of pieces of content every *hour*.[6] This is a fraction of the content actually posted to their sites. Even the largest bureaucracy of content moderators could not possibly hope to contain harmful content on their services if they attempted to do so only by reviewing individual posts one at a time.

Since its beginnings, behavioral and associational moderation has greatly expanded. Twitter has rolled out a new "coordinated harmful activity" policy,[7] following the need to find a policy to respond to situations like QAnon adherents harassing Chrissy Teigen en masse.[8] Meta has new policies against "coordinated social harm,"[9] "brigading" and

"mass reporting."[10] No platform has ever given much of an explanation of what constitutes "coordination." When does a deluge of tweets become coordinated harassment? How does a platform determine if an individual user is coordinating with others? When does the way a group acts online or offline create "social harm"?

In other words: In the eyes of these platforms, when does a crowd become a mob? We simply don't know. Platforms have taken down tens of thousands of accounts under these umbrellas with little explanation.[11]

To date, enforcement of these policies has been patchy and opaque. Meta has publicly invoked its "coordinated social harm" *once*, seemingly at random, to justify a take-down of a network associated with a movement in Germany called "Querdenken" that spread conspiracy theories about COVID-19 and had been tied to off-platform violence.[12] No further details were given. It's unclear if Twitter still enforces its coordinated harmful activity policy, or whether it even looks to enforce it proactively. The policy technically sits under Twitter's rules about platform integrity, but Twitter only reports on subsets of its enforcement under those policies. After a co-author and I argued that the opacity of these distinctions was problematic, Twitter expanded its previously sui generis data transparency for information operations to these categories as well, acknowledging that there is no good reason for them to be treated differently.[13] Other platforms are yet to follow for their equivalent policies.

In the eyes of platforms, when does a crowd become a mob? We simply don't know.

Yet, while there are increased calls for transparency from platforms more generally, transparency deficits in the realm of behavioral and associational content moderation go relatively unremarked upon. And for all the talk about threats to freedom of speech or expression online, threats to freedom of association are rarely mentioned.

But freedom of association is an equally important part of a vibrant and healthy public sphere. The ability to act in groups is vital and worth protecting because, as the Supreme Court has said, "[e]ffective advocacy of both public and private points of view, particularly controversial ones, is undeniably enhanced by group association."[14] And beyond these collective goods, "freedom to engage in association for the advancement of beliefs and ideas is an inseparable aspect" of individual liberty.[15] Of course, as with freedom of speech, no right is absolute and user association can and should be the subject of moderation—for all the reasons discussed above. But the democratic and individual interests at stake in interfering with people's ability to associate means these decisions deserve careful scrutiny. Instead, despite freedom of association being increasingly implicated in content moderation actions, it gets far less attention than other forms of platform intervention.

Transparency for transparency's sake is not useful, and so it's important to be clear on the unique reasons great platform transparency is important in this context. First, it can ensure that these ostensibly non-content-based content moderation decisions are, well, actually non-content based. Without understanding more about how platforms delineate the borders of groups or the kinds of crowd-like behaviors they designate as harmful, it is impossible to know for sure that these supposedly content-neutral tools are not being deployed in selective ways based on criteria related to the substantive content involved. Why, for example, did Meta target Querdenken and, as far as anyone can tell, no one else? Was Twitter's new coordinated harmful

> As regulatory proposals proliferate, few recognize the existence of associational moderation.

activity mainly a figleaf to give them a way to try keep Crissy Teigen on the platform? When and how are platforms deciding to act on the basis of these policies?

Second, understanding how policies are enforced will help minimize collateral damage. Are innocent users, with no particular connection to the coordinated activities platforms have deemed harmful, caught up in these broad platform removals? When a platform takes down tens of thousands of QAnon accounts, are they all equally liable for harmful behavior, or are some users being found guilty by association with a few prominent accounts? Are platform responses proportionate to the actual behavior of an individual in question, or are all accounts in a group subject to the same sanction regardless of the level of their participation? There is currently no way to answer these questions.

Third, how are platforms collaborating in sharing information about group dynamics? Adversarial groups rarely target a single platform alone, and will often coordinate across platforms, leveraging their different affordances in order to maximize impact. Platform collaboration is therefore likely to be essential to effective moderation of groups.[16] But it also risks breaches of user trust in sharing information in a manner inconsistent with user expectations, and could compound the effects of mistakes if platforms make a uniform error as a result of acting collectively.

There are also unique costs to transparency in the behavioral moderation context though. If platforms are too explicit about the behavioral signals they use to find problematic coordination, this will enable adversaries to better conceal their tracks.[17] There are thus difficult and unique trade-offs involved in bringing accountability to behavioral moderation in a way that protects associational rights without enabling further harmful behavior by bad actors.

The first step in solving these problems is for policymakers to acknowledge that they exist. As regulatory proposals proliferate globally, few even recognize the existence of group or associational moderation. Most focus on regulation of post-by-post content moderation decisions, and introduce transparency and due process requirements individual cases. These mandates will do little to shed light on group moderation. Worse than that, when transparency focuses on one kind of platform decision to the exclusion of others, it risks pushing more controversial and difficult decisions into the shadows.

Crowd-based moderation is an important tool in the content moderation toolbox. It allows platforms to respond to harms on their services in a way that is different in nature and scope to moderation of individual posts. But unless accountability systems for this kind of content moderation are built in from the start, we risk accountability deficits that are exactly the same in nature as we have seen in the moderation of individual posts: inaccurate, uneven, and biased enforcement, and a lack of investment by platforms in fixing these problems.

Notes

1 *See* Camille François & Evelyn Douek, *The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations*, 1 J. Online Tr. & Safety 1, 7 (2021).

2 *See* Cecilia Kang et al., *Russia-Financed Ad Linked Clinton and Satan*, N.Y. Times (Nov. 1, 2017), https://www.nytimes.com/2017/11/01/us/politics/facebook-google-twitter-russian-interference-hearings.html.

3 Nathaniel Gleicher, *How We Respond to Inauthentic Behavior on Our Platforms: Policy Update*, Facebook Newsroom (Oct. 21, 2019) [hereinafter Gleicher, *Inauthentic Behavior*], https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update/ (emphasis in original).

4 Gleicher, *Inauthentic Behavior*, *supra* note 3.

5 *See* Evelyn Douek, *The Free Speech Blind Spot: Foreign Election Interference on Social Media*, *in Defending Democracies: Combating Foreign Election Interference in a Digital Age* 265 (Jens David Ohlin & Duncan B. Hollis eds., 2021).

6 *See* Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 526, 537–38 & nn.38–40 (2022).

7 *See Coordinated Harmful Activity*, Twitter Rules (last visited Apr. 2, 2021), https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity.

8 *See* Evelyn Douek, *Twitter Brings Down the Banhammer on QAnon*, Lawfare (July 24, 2020), https://www.lawfareblog.com/twitter-brings-down-banhammer-qanon.

9 *See* Nathaniel Gleicher, *Removing New Types of Harmful Networks*, Facebook Newsroom (Sept. 16, 2021) [hereinafter Gleicher, *Harmful Networks*], https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/.

10 *See* Gleicher, *Harmful Networks*, *supra* note 9; Nathaniel Gleicher et al., Meta, *Adversarial Threat Report* (2021), https://about.fb.com/wp-content/uploads/2021/12/Metas-Adversarial-Threat-Report.pdf.

11 *See An Update to How We Address Movements and Organizations Tied to Violence*, Facebook Newsroom (Aug. 19, 2020), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/; Douek, *supra* note 8.

12 *See* Gleicher, *Harmful Networks*, *supra* note 9; Jeff Horwitz & Justin Scheck, *Facebook Increasingly Suppresses Political Movements It Deems Dangerous*, Wall St. J. (Oct. 22, 2021), https://www.wsj.com/articles/facebook-suppresses-political-movements-patriot-party-11634937358.

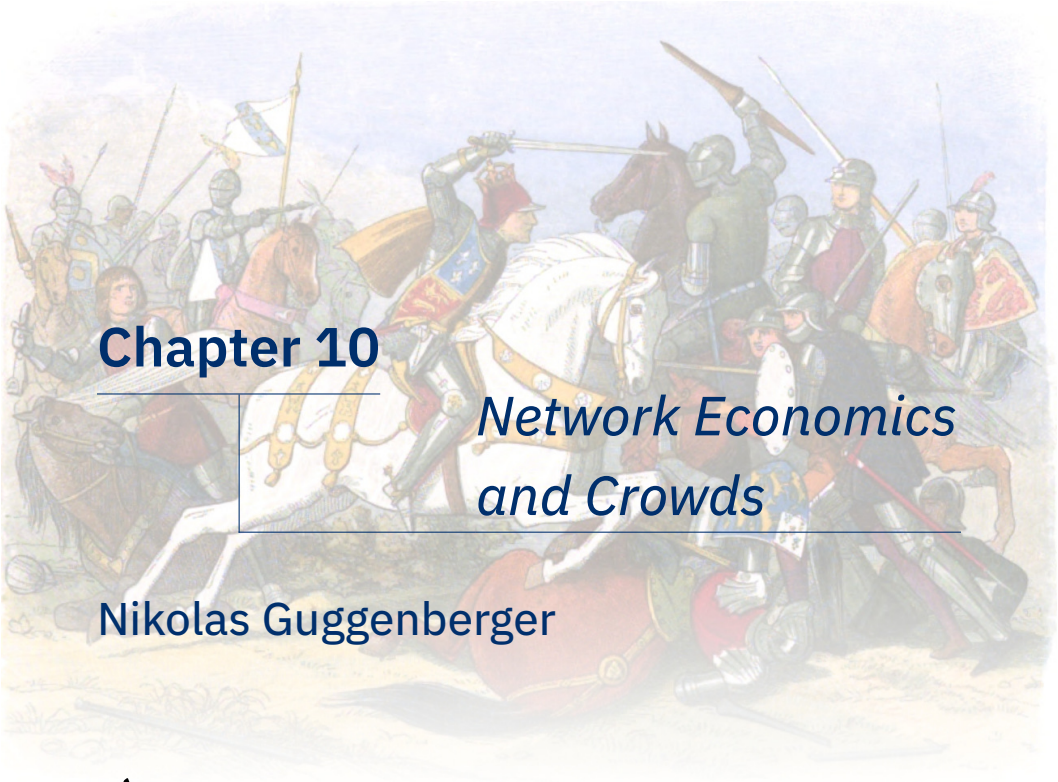
13 *See* François & Douek, *supra* note 1; Yoel Roth & Vijaya Gadde, *Expanding Access Beyond Information Operations*, Twitter Blog (Dec. 2, 2021), https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-.

14 *See NAACP v. Ala. ex rel. Patterson*, 357 U.S. 449 (1958).

15 *See Patterson*, 357 U.S. 449.

16 *See* Evelyn Douek, *The Rise of Content Cartels* (Knight First Amend. Inst. at Columbia Univ. 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels.

17 *See* Google, *How Google Fights Disinformation* 3 (Feb. 2019), https://blog.google/documents/37/How_Google_Fights_Disinformation.pdf ("[W]e try to be clear and predictable in our efforts, letting users and content creators decide for themselves whether we are operating fairly. Of course, this is a delicate balance, as sharing too much of the granular details of how our algorithms and processes work would make it easier for bad actors to exploit them."); Sara Harrison, *Twitter's Disinformation Data Dumps Are Helpful—to a Point*, Wired (July 7, 2019), https://wired.com/story/twitters-disinformation-data-dumps-helpful/, ("Twitter would not reveal any specifics about its process for this article. 'We seek to protect the integrity of our efforts and avoid giving bad actors too much information, but in general, we focus on conduct, rather than content.' ").



Chapter 10

Network Economics and Crowds

Nikolas Guggenberger

A handful of digital platforms dominates the social media landscape. One company stands out: Meta, with Facebook, Instagram, and WhatsApp. Depending on the definition of social media there are also TikTok, Twitter, YouTube, and Twitch. None of these services are perfect substitutes for each other, however. They serve different purposes. That leaves users with little choice and hardly any realistic exit options. Users would need to leave many of their social connections behind. Not least due to lack of credible threats to exit the platforms, users also lack internal voice.[1]

In reaction to the enormous market concentration among social media platforms and the seemingly never-ending revelations of abusive behaviors, antitrust remedies and competition policy reforms have gained attention in the U.S. and abroad. The proposals range from changes to the corporate structure and forced divestitures[2] to tougher merger rules,[3] behavioral requirements or limitations, including interoperability regimes and duties to deal,[4] and technical standards.[5]

Many, including the authors of the influential 2020 House Majority Report, favor combinations of these measures.[6] While the views on the justification of antitrust remedies differ, the basic idea behind reining in monopolies is that economic concentration among digital platforms threatens democratic discourse, stifles innovation, and causes harm to individuals.[7]

The concerns about democratic discourse are rooted in the concentration of political power and the creation of single points of failure in the communication architecture.[8] If a single entity—or founder CEO as is the case at Meta due to its centralized internal governance structure—controls the communicative environments of billions of people, that entity wields significant sway over the political decision-making process and elections.[9] Mistakes can lead to systemic repercussions, even if the entity were to hold only the best intentions. On a technical level, single points of failure can increase the likelihood of systemic outages, something the internet's decentralized architecture all but entirely precludes.

The logic behind concerns about individual harm is straightforward, textbook-like industrial organization. In a competitive market, users have alternatives. If they do not like the services provided by one of the social media platforms, they can choose a different one. The re-

At the core of the relationship between barons and the crowds lies a barter. Market concentration affects both ends of this barter.

sulting competition drives prices down and quality up. It also spurs innovation. With increasing market concentration, that changes. Monopolists can charge higher prices or provide lower quality, while the output of the entire industry decreases. They do that to increase their profits. These basic economic mechanisms apply to social media as they do to car manufacturers or banks.

At the core of the relationship between barons and the crowds lies an intransparent barter. Social media platforms provide social media services. These services include connectivity, information organization and content moderation. Social media users, from consumers to influencers and politicians, provision their data, their attention, their content, and their presence as nodes in a network.[10] All of that is valuable to platforms. Some users engage in that barter deliberately; most, however, likely do not realize the full extent of the exchange. Either way, this barter describes the economic reality of the transaction between platforms and users.

Market concentration affects both ends of this barter. The dominant advertisement-based funding model with its monetary prices of zero for users neither mitigates the impact of market concentration on the economic exchange, nor does it preclude antitrust remedies.[11] Instead of charging higher prices, monopoly platforms can simply lower the quality of their services by reducing the investments in content moderation below what they would need to offer in a competitive environment, for example.[12] They might also deteriorate users' privacy protections or expose them to more advertising—all compared to hypothetically less concentrated markets.

Standing antitrust doctrine, however, is not well equipped to mitigate the threats from excessive concentration among digital platforms. It conditions interventions on levels of market concentration exceeding what threatens democratic discourse or may harm individuals. As

constructed and applied by courts, antitrust law further requires specific anticompetitive behavior and economically quantifiable harm to the transactions' counterparts. That can be hard to prove in court, especially with the standards of proof required in recent case law.[13]

The various reform proposals aiming at reining in the concentration of power among social media platforms fall into at least three different categories. First there are structural reforms, named after their immediate impact on the structure of the firms operating social media platforms and the market in which they operate. Structural reforms may include horizontal break-ups or functional separations. Both have a long tradition in U.S. antitrust and regulation. The former draw on the historic example of the Standard Oil breakup,[14] the latter build on New Deal–type banking regulation, *à la* the Glass–Steagall Act, and similar arrangements in the telecommunications and railroad sector.[15]

Horizontal breakups would, for example, split up Facebook Blue into several mini-Facebooks or undo past mergers, like then-Facebook's acquisitions of Instagram and WhatsApp, as demanded by the FTC in its lawsuit against Meta.[16] These approaches aim at reestablishing competition by converting one big player into several smaller players. In a sense, this remedy would reset the market and competition to an earlier stage. While certainly the most immediate relief to a monopoly problem, some have articulated concern that the remedy might restart a period of competition for the market, eventually again culminating in market concentration.[17] Such criticism, however, assumes that the remedy is pursued in isolation. Complementary procompetitive reforms could prevent a reverting to pre-remedy structures.

Before her tenure as FTC Chair, Lina Khan championed a form of vertical breakup, the "separation of commerce and platforms."[18] This remedy draws different lines: it isolates functions and aims to resolve

conflicts of interest inherent in these functions. Applied to app stores, for example, it would prevent app store operators from also offering their own apps on the platform—a practice common among both the Apple iOS Store and the Google Play Store. It may also alleviate some concerns inherently linked to size and power, as it would limit the reach and confine the lines of business of digital platforms. It could also, albeit indirectly, reshape the environment for mergers and start-ups. After all, integrated Big Tech platforms are among the most active acquirers of new businesses and define the exit opportunities for start-ups.[19]

The implications of breakups for privacy are contested. Revived competition could provide users with more privacy-sensitive alternatives. Some argue, however, that breakups would distribute personal data among several independent entities which either in itself constitutes a privacy harm or leaves that data more vulnerable. Understanding data and privacy as relational constructs[20] offers yet another perspective on the potential impacts of breakups on privacy.

While it appears clear that discourse would certainly evolve in reaction to breakups, the details likely depend on the framework accompanying any future breakups. Openness or interoperability between the newly created competitors would bring entirely different conse-

> Without simultaneous and mandatory interconnection, discussions might become more isolated, reinforcing biases and extremism.

quences than insular small platforms. At a very basic level, the former would extend individuals' reach, while the latter has the potential to sever connections or, at least, insert significant friction in communicative processes and social graphs. Without simultaneous and mandatory interconnection, discussions might become more isolated, playing into reinforcing biases and extremism in discourse.

Next, some regulators, policymakers, and scholars have focused on opening up platforms as complementary or stand-alone policies to strengthen competition.[21] Core to these approaches is lowering switching costs for users and reducing market entry barriers for nascent competitors. There are two different types of approaches to easing migration between platforms: data portability provisions and interoperability frameworks. Data portability provisions are meant to enable transitioning existing data sets from one platform to another. They might apply broadly to personal data, such as article 20 GDPR, or cover a specific subset of the data stored or created by the digital platform. In theory, users willing to switch their provider can take their profiles and personal data with them and start on a competitor's platform right where they left off. In practice, data portability provisions have proven rather toothless. Technical difficulties with the data formats, incompatible platform features, and the contextual nature of data have impeded the effectiveness of this regulatory tool. Drawing from the structures of telecommunication networks or open email protocols,[22] respectively, more recent commentary and initiatives have focused on interoperability, that is, establishing connections beyond the boundaries of the platforms' own network.[23] Practically speaking interoperability would allow users to send messages from Facebook to Twitter or from WhatsApp to Signal, for example. Hopes are that interoperability more profoundly revives competition than data portability rules could.

While data portability between platforms can easily fit into consent based privacy frameworks, interoperability requirements might prove less compatible. Depending on the specificities of the interoperability requirements, the exchange of messages, posts, pictures, and more beyond the boundaries of a closed network requires the sharing of personal data with third parties, namely the receiving network. Double opt-in frameworks, on the senders' and the receivers' end, can likely overcome these concerns, but might weaken the interoperability framework's effectiveness in reviving competition. Paradigm shifts in privacy protections away from consent and notice regimes and toward democratic control[24] or fiduciary responsibilities[25] could alleviate the tensions between the current data protection frameworks and meaningful competition reforms.

Access rights, whether rooted in antitrust doctrine in the form of the essential facilities doctrine, regulation, or legislation, usually take the form of a duty to deal.[26] They inevitably include elements of interoperability. Other than the previously discussed interoperability frameworks, however, they generally aim at giving downstream market participants access to the platform, not at connecting competing platforms in a horizontal manner. In practice, that could mean giving App developers rights against the App stores.

Access rights may, but do not have to be combined with prohibitions of self-preferencing. Self-preferencing is a practice describing platforms granting favorable treatment to their own services offered on the platform compared to third-party services. Apple, for example, is said to have manipulated the rankings of apps in the iOS App store in favor of its own music service. Several bills under consideration address self-preferencing.

Some criticize that horizontal interoperability requirements and, especially, access rights with self-preferencing prohibitions impede

platforms' abilities to moderate content.[27] The concern is that the app stores, for example, might be prevented from banning the likes of Parler from its platforms, exacerbating disinformation online.[28] Yet, as recent scholarship has shown, these concerns are overblown. In fact, reasonable deplatforming has long been the norm for common carriers.[29] Also, large platforms are not the only entities that could provide content moderation services.[30]

Applying existing antitrust laws is just one way to implement some of the currently discussed remedies. Competition could also be strengthened via pro-competitive legislative reforms, either in the form of industry specific competition policy or more foundational reforms to the antitrust statutes. Antitrust, pro-competitive regulatory regimes, and regulation in general do not preclude each other.[31] In fact, they often make for necessary complements. Even *Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*, the Supreme Court's gutting of the essential facilities doctrine, only established an assumption about Congressional intent: when Congress regulates, it allegedly implies that it has done so exhaustively.[32] When defining the future competitive environment *Trinko* does not hinder a symbiosis of regulation of antitrust and regulation; at most, it requires an explicit legislative affirmation.

To summarize, social media markets share many characteristics with other concentrated markets. Its monopolists will try to extract monopoly rents just as other companies would. The difference lies in the types of harm social media behemoths can cause compared to other industrial giants. Plenty of pro-competitive remedies are under consideration at the moment. Some have better chance of becoming law than others. All, however, have the potential to change the dynamics between barons and the mob fundamentally.

Notes

1 On exit and voice as alternative and complementary methods of exercising influence, see Albert O. Hirschman, *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States* (Harvard Univ. Press 2004); Nikolas Guggenberger, *Moderating Monopolies*, 38 Berkeley Tech. L.J. 119 (2023) [hereinafter Guggenberger, *Moderating*].

2 *See* First Amended Complaint, *FTC v. Facebook*, No. 1:20-cv-03590, (D.D.C. Aug. 19, 2021); Zephyr Teachout, *Break 'Em Up: Recovering Our Freedom from Big Ag, Big Tech, and Big Money* (2020); Lina M. Khan, *The Separation of Platforms and Commerce*, 119 Colum. L. Rev. 973 (2019), https://columbialawreview.org/wp-content/uploads/2019/05/Khan-THE_SEPARATION_OF_PLATFORMS_AND_COMMERCE-1.pdf; Rory Van Loo, *In Defense of Breakups: Administering a "Radical" Remedy*, 105 Cornell L. Rev. 1955 (2020).

3 *See* C. Scott Hemphill & Tim Wu, *Nascent Competitors*, 168 U. Pa. L. Rev. 1879 (2020).

4 *See* Zachary Abrahamson, *Essential Data*, 124 Yale L.J. 867 (2014); Nikolas Guggenberger, *Essential Platforms*, 24 Stan. Tech. L. Rev. 237 (2020) [hereinafter Guggenberger, *Essential*], https://law.stanford.edu/wp-content/uploads/2021/05/publish_this_-_guggenberger_essential_platforms_eic.pdf.

5 *See* Mike Masnick, *Protocols, Not Platforms: A Technological Approach to Free Speech* (Knight First Amend. Inst. at Columbia Univ. 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

6 *See* Subcomm. on Antitrust, Commerce & Admin. L. of the House Comm. on the Judiciary, *Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations* (Oct. 2020); Lina M. Khan, *Amazon's Antitrust Paradox*, 126 Yale L.J. 710, 790–801 (2017).

7 *See* Guggenberger, *Moderating*, *supra* note 1.

8 *See* Subcomm. on Antitrust, Commerce & Admin. L. of the House Comm. on the Judiciary, *supra* note 6, at 1–2.

9 *See* Guggenberger, *Moderating*, *supra* note 1.

10 *See* Guggenberger, *Moderating*, *supra* note 1.

11 *See* John M. Newman, *Antitrust in Zero-Price Markets: Foundations*, 164 U. Pa. L. Rev. 149 (2015); John M. Newman, *Antitrust in Zero-Price Markets: Applications*, 94 Wash. U. L. Rev. 49 (2016).

12 *See* Marshall Steinbaum, *Establishing Market and Monopoly Power in Tech Platform Antitrust Cases*, 67 Antitrust Bulletin 1, 9 (2022).

13 *See Ohio v. Am. Express Co.*, 138 S. Ct. 2274 (2018).

14 *See Standard Oil Co. v. United States*, 221 U.S. 1 (1910).

15 *See* Julia Maues, *Banking Act of 1933 (Glass-Steagall)*, Fed. Rsrv. Hist. (Nov. 22, 2013), https://www.federalreservehistory.org/essays/glass-steagall-act.

16 *See* First Amended Complaint, *FTC v. Facebook*, No. 1:20-cv-03590, (D.D.C. Aug. 19, 2021).

17 *See* Francis Fukuyama et al., *Report of the Working Group on Platform Scale* (2020), https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale.

18 *See* Khan, *supra* note 2.

19 *See* Mark A. Lemley & Andrew McCreary, *Exit Strategy*, 101 B.U. L. Rev. 1 (2021).

20 *See* Salome Viljoen, *A Relational Theory of Data Governance*, 131 Yale L.J. 573 (2021), https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance.

21 *See* Michael Kades & Fiona Scott Morton, *Interoperability as a Competition Remedy for Digital Networks* (Wash. Ctr. for Equitable Growth, working paper, Sept. 2020), https://equitablegrowth.org/working-papers/interoperability-as-a-competition-remedy-for-digital-networks/; Nikolas Guggenberger, *Essential Platform Monopolies: Open Up, Then Undo*, Pro Mkt. (Dec. 7, 2020), https://promarket.org/2020/12/07/essential-facilities-regulation-platform-monopolies-google-apple-facebook/.

22 *See* Masnick, *supra* note 5.

23 *See* Kades & Scott Morton, *supra* note 21; Guggenberger, *Essential*, *supra* note 4; Cory Doctorow, *Adversarial Interoperability: Reviving an Elegant Weapon From a More Civilized Age to Slay Today's Monopolies*, Elec. Frontier Found. (June 7, 2019), https://www.eff.org/deeplinks/2019/06/adversarial-interoperability-reviving-elegant-weapon-more-civilized-age-slay.

24 *See* Viljoen, *supra* note 20.

25 *See* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 UC Davis L. Rev. 1185 (2016).

26 *See* Guggenberger, *Essential*, *supra* note 4.

27 *See* Jane Bambauer & Anupam Chander, *Bills Meant to Check Big Tech's Power Could Lead to More Disinformation*, Wash. Post (June 6, 2022), https://www.washingtonpost.com/outlook/2022/06/06/antitrust-bills-big-tech-hate-speech-disinformation/.

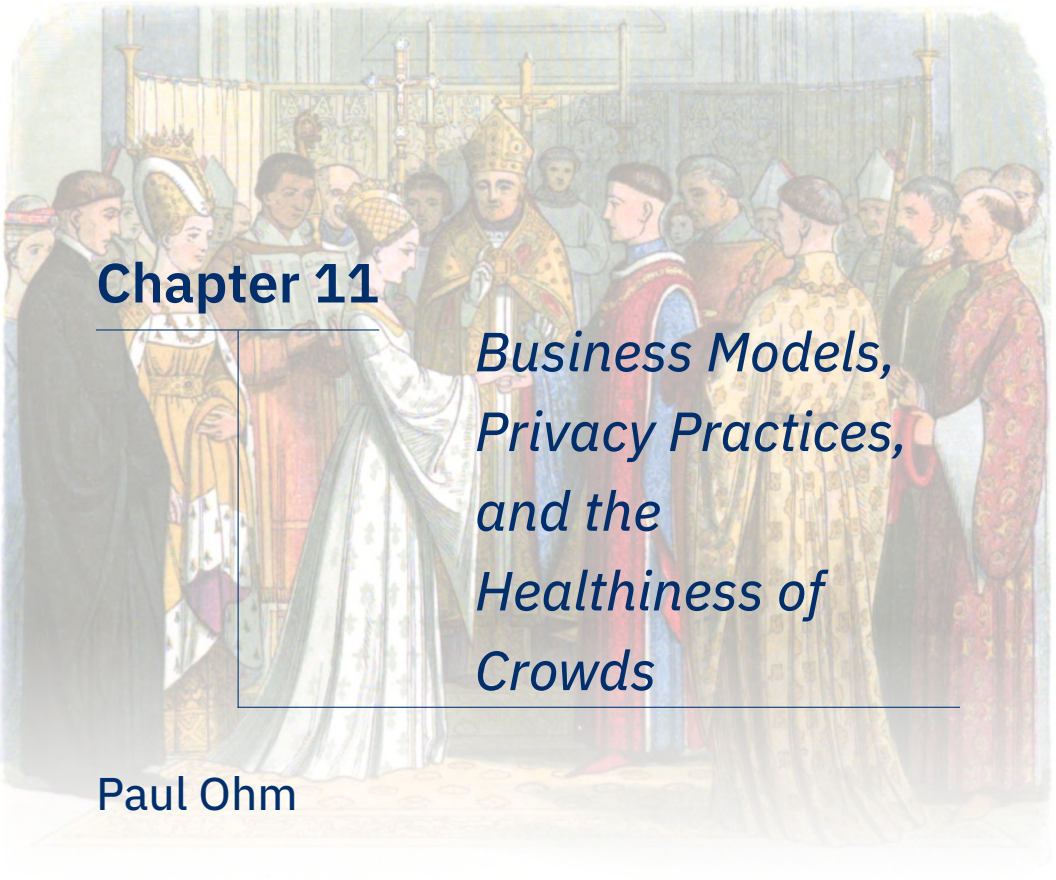
28 *See* Bambauer & Chander, *supra* note 27.

29 *See* Ganesh Sitaraman, *Deplatforming*, 133 Yale L.J. 419 (2023).

30 *See* Fukuyama et al., *supra* note 17.

31 *See* Elettra Bietti, *Structuring Digital Platform Markets: Antitrust and Unfair Competition*, U. Ill. L. Rev. (2024).

32 *See Verizon Commc'ns Inc. v. Law Offs. of Curtis V. Trinko, LLP*, 540 U.S. 398 (2004).



Chapter 11

Business Models, Privacy Practices, and the Healthiness of Crowds

Paul Ohm

The crowds and mobs we are focused on in this volume assemble in the virtual spaces provided by various online services. Each service offers its own mix of technical affordances, social norms, and governance structures, and these together play a lead role in shaping the types of communities they tend to attract. By comparing these features across different online services, we might learn how to design or redesign services to encourage healthy crowds and discourage unhealthy mobs.

Focus in particular on the way the intertwined "policy knot"[1] of business models, privacy commitments (or obligations), and technological design seems to influence the overall health and stability of dif-

ferent crowd-generating online services. My hypothesis is that services built on business models that do not require the collection of massive amounts of user data tend to lead to healthier spaces, meaning spaces less prone to various problems including misinformation and disinformation, harassment, trolling, and other forms of toxicity, than services built on business models requiring the tracking of behavior. If so, it would give us a roadmap for redesigning toxic services and maybe even serve as a powerful justification to discourage or outlaw surveillance-based business models.

The baseline for the comparisons that follow are the massive social networking services supported by behavioral advertising models: Facebook, Instagram, Twitter, and TikTok; call them "traditional social media." These services are supported by *behavioral advertising* business models, which require the extraction and exploitation of private information about users in order to sell individually targeted ads. In contrast, I offer two case studies built on a very different model: podcasts and Reddit. Both earn revenue primarily (though not exclusively) through *contextual advertising*—advertisers pay for access to particular audiences or communities, organized around broadly drawn interests or identities, rather than historical records of the specific behav-

My hypothesis is that services built on business models that do not require the collection of massive amounts of user data tend to lead to healthier spaces.

ior of individuals. I contend that Reddit and podcasts tend to support healthier crowd dynamics than those found in traditional social media, which may be a result of liberating them from the incentives imposed on their surveillance focused counterparts.[2]

Podcasts Podcasts make money primarily through in-stream audio advertisements. The podcaster sells and the advertiser buys an audience, as in the olden days of pre-Internet television, radio, or newspaper advertising. The advertiser's willingness to pay depends on its predictions about the consumer behavior of the group-qua-group, such as "people who like true crime stories tend to buy home security systems," or "people who follow this host tend to buy the products she personally endorses." The podcaster receives no advertising benefit from knowing anything additional about any individual listener, because there is no way to play a tailored ad, just for them.

Why are podcasts built on this business model? Blame (or credit) technical affordances. Until recently, most podcasts were delivered exclusively via RSS, a distributed and decentralized publishing format most commonly associated with the heyday of blogging.[3] With RSS, podcasters publish episodes as separate, downloadable files on the public web, using a standardized and open format called an RSS feed.[4] Anybody can then develop a tool—a "feed reader" in the case of blogs and sometimes called a "podcatcher" for podcasts—capable of downloading these files to a computer or smartphone for later consumption. Because of the built-in limitations of RSS, the only thing the blogger or podcaster can learn about each download is that a device from a particular IP address, bearing a particular user-agent string (which can reveal the type and version of podcatcher and operating system used but not much more), downloaded a particular episode on a particular date and time.[5] Podcasters usually don't have access to the identity of

the user downloading the episode, much less the rich stream of other behavioral data available to traditional social media.

Stories suggest that podcasting is built upon such a minimalist, arguably outdated, technological architecture and business model because Apple saw itself more as a platform for podcasters rather than a partner in the monetization of podcasts, so it failed to add features to iTunes that could have provided behavioral information to podcasters.[6] If it had wanted to, Apple could have embraced a streaming model for podcasts, one in which the podcaster would have received far more information than available via RSS, such as analytics about whether and when episodes had been played, finished, and replayed.

The contextual advertising model still generates healthy revenue for many podcasters. One study suggests the entire podcasting industry generated more than a billion dollars in 2021.[7]

Reddit Reddit is a discussion-oriented service organized around specific message boards called "communities" or "subreddits". Each subreddit focuses on a particular topic, say a hobby like r/DIY, a fandom like r/strangerthings, or an interest like r/movies. Some of the most popular subreddits feature content that delivers a particular emotional experience, such as r/funny or r/aww, or that shares a specific approach to learning, such as r/todayilearned or r/explainlikeimfive. Subreddits run the gamut from massive communities of millions of users to niche communities of only a few.[8] Each subreddit is created by a user who becomes the group's moderator, has the power to remove posts and appoint other moderators, and typically promulgates subreddit-specific moderation rules.

Most of Reddit's revenue comes from advertising targeted at the subreddit rather than individual user level. Advertisers often specify subreddits specifically by name. Alternatively, advertisers might tar-

get specific topics or keywords, which Reddit's ad platform translates into particular subreddits. Reddit generated $350 million in revenue in 2021[9] and has been valued at more than $10 billion.[10]

Why These Affordances Might Produce Healthier Communities

My claim is that podcasts and Reddit tend to create healthier and less toxic communities than traditional social media services. I can't yet prove that this is true, but I offer four reasons to support the claim: The contextual business model incentivizes podcasters and Reddit to nurture healthy communities; these incentives result in the creation and maintenance of legible boundaries between sub-communities; legible boundaries and delegated moderation increase the tools available to combat unhealthy group dynamics; and all of these factors increase the trust users have in the administrators of a service.

Focusing on incentives Consider first the incentives that are placed on the providers of services, and the way the contextual advertising model might incline Reddit and podcasters to put more emphasis on nurturing healthier communities.

To generate revenue, Reddit and podcasters need to not only grow their audience but also attract an audience with a clear identity they can use to attract certain kinds of advertisers. Fringe subreddit moderators or podcasters might develop a destructive, unhealthy community in order to attract certain advertisers—surely someone wants to market their products or services to misogynists, vaccine deniers, or racists. To attract a much larger set of advertisers, and a much larger pool of advertising budgets, however, subreddit moderators and podcasters will nurture healthy communities, those with norms (or rules) against harassment, toxicity, or misinformation.

Importantly, podcasters and Reddit both have no revenue-fueled incentive to segment their audience into individuals; they are encouraged to treat them as a collective, a group or community. This might be a positive or negative force. Treating someone as a member of a group rather than an individual might feed moblike behavior, by encouraging people to get swept up in negative group dynamics. On the other hand, it might help members avoid feelings of isolation or alienation.

Legible boundaries Podcasters and Reddit tend to create distinct communities with unmistakably clear and legible boundaries. A listener or user knows when they leave one podcast to listen to another, or read a post in one subreddit rather than another. To be clear, the UI of most podcast listening apps and Reddit commingle content from across podcasts and subreddits, respectively, mimicking the "infinite scroll" design pattern of traditional social media, but they use prominent labels and other visual cues to indicate the source—the specific podcast or subreddit—of each piece of content in the scroll.

Legibility is a product of the contextual advertising model. Advertisers will pay a premium for one subreddit or one podcast over another, so creating legible boundaries is in the economic self-interest of the providers.

For users, legibility serves a beneficial purpose delineating the boundaries of different communities. Each community develops its own ethos, norms, culture, and social context. Users can exit communities that develop undesirable norms–such as subreddits that become toxic.

Legible communities thus give users agency. A subreddit user chooses whether to join or not join a particular subreddit. Ads shown in a subreddit are linked in the user's understanding to this intentional choice they made in the past.

Contrast this legibility with the well-documented illegibility on traditional social media services. These services blur the distinctions between advertising and "organic" content, leading some critics to allege consumer confusion and manipulation. Of all of them, TikTok has most famously devised an entire design that blurs the lines between different interests, groups, and communities.[11] Its infinite scroll, and the removal of the requirement to even click on a link to engage with content, mushes all TikTok content into one disaggregated stream, tailored specifically for each individual.

Tools for combatting harms These incentives and legible boundaries give podcasters and Reddit's management and moderators tools to combat toxicity, misinformation, or harassment that traditional social media lacks. First, users can exit a community without leaving the entire service. A simple click is all that is required, and the user remains a member of all of their other communities. The wound can be cauterized cleanly, quickly, and with no uncertainty. This mechanism also allows for "voting with ones feet," giving users the threat of mass exodus as a means to ensure their concerns and misconduct reports are taken seriously.

Second, pick your pandemic metaphor. The impermeability of boundaries between communities limits contagion from one community to others. A hard boundary between communities can help quarantine a sick community from healthier adjacent ones. Traditional social media lack these quarantine boundaries, making them more like a single, massive room teeming with users, more prone to the viral spread of harmful speech or conduct across their services.

Third, unlike the flatness of the governance of traditional social networking sites—a simple binary system between "the platform" and the disaggregated "users"—subreddits and podcasts develop gover-

nance hierarchies. Subreddits are led by moderators and individual podcasts are associated with a particular individual or specific team. These official leaders can announce rules, facilitate (or not) democratic participation, shape the allowable discourse in their subreddit or podcast, and punish or exile those who break the rules.

These hierarchies provide new and more powerful methods of content moderation. Rather than face the challenge of providing global content moderation for hundreds of millions of users, Reddit's management can treat its army of moderators as a front line force for moderation, each well-tuned to different contexts and subcultures on the site. Some moderators have no problem handling the content moderation challenges in their subreddits with no need for help from Reddit management. This frees management to focus more on ungoverned or poorly governed subreddits, narrowing the content moderation challenge.

This also support a distributed form of governance, allowing individual subreddits or podcasts to experiment with different approaches, giving rise to a "laboratory of content moderation." Good ideas can be copied by other moderators or podcasters, and bad ideas will fail to propagate.

Trust All of these forces increase the odds that users will trust a particular podcast, a particular subreddit, or a company such as Reddit. Some have tied the idea of trust to user knowledge and consent.[12] The podcast and Reddit model bake consent in. A user chooses to enter or exit a specific community, and there is a premium placed on making that choice clear and easy to manage. Imagine the frustration a user would feel if Reddit or a podcaster did not respect this choice–if joining one subreddit auto-joined them into another, or if a dark pattern tricked a user into subscribing to a podcast they didn't want. There is

a premium placed on clear, unambiguous messaging and mechanisms about entry and exit.

Legible boundaries help users properly direct both praise and blame where they are due. A user might understand that a particular moderator has let a particular subreddit turn into a toxic mess, increasing their admiration for and trust in the moderators of subreddits they continue to value.

The enhanced trust that comes from these mechanisms might even strengthen the business model. An anecdote: I have never in my 30 years using the commercial Internet willingly clicked on an ad while using a service that I suspect or know uses behavioral advertising. I am wary to click on such ads, for fear that doing so would brand me an "easy mark," one worth targeting with future ads. In contrast, I have often purchased items advertised by the podcasters I follow. I know that my support is not feeding the surveillance economy. Buying from these advertisers gives me an easy and relatively risk-free way to support content I value.

Storm Clouds in the Distance

To be clear, this does not mean that podcasting and Reddit have been free from problems. Reddit was once a notorious site for subreddits hosting revenge porn and naked photos of celebrities.[13] It has also housed subreddits that were cesspools of misinformation, harassment.[14] Although many such subreddits have since been banned, controversial subreddits remain on the site.[15]

Podcasts are arguably even more of a mixed bag. Many popular podcasters are known to promulgate problematic content and attract toxic followings. The most listened-to podcaster in the world, Joe Rogan, has frequently spread misinformation about vaccines,[16] and has

repeatedly said offensive and demeaning things.[17] Logan Paul, another popular but controversial figure, has criticized a "vocal minority" of his own audience for their toxicity.[18] Problematic content isn't restricted to the most popular podcasts. The democratic nature of RSS means that anyone can start a podcast, with few gatekeepers standing in the way of problematic content.[19]

In addition, although I am arguing for the benefits of contextual over behavioral advertising models, this is not to pine for the "good old days" of broadcast television, when powerful gatekeepers controlled the culture, perpetuated archaic values, and excluded most voices. Contextual advertising of the kind represented by podcasting and Reddit is less subject to the narrowness and conservativism of the pre-Internet age, permitting much more democratic, decentralized, and eclectic fora for content and community today.

In a moment, I will identify what traditional social media should learn from podcasting and Reddit and how they ought to shift from behavioral to contextual advertising models, voluntarily or by regulation; unfortunately, there are pressures working in the opposite direction, pushing both podcasting and Reddit toward a behavioral advertising model. Many key actors in the podcasting economy are trying to "bolt on" surveillance atop RSS to learn more about podcast user behavior.[20] In 2014, a company called Acast devised "dynamic ad insertion," which tailors the ads bundled in a podcast to information known

> There is a premium placed on clear, unambiguous messaging and mechanisms about entry and exit.

about the user at the time of download.[21] This is far from the just-in-time hyper-personalization of the web, but it breaks the purely contextual, one-community model by letting different listeners hear different ads for the same episode, ads tailored to their prior behavior.

This might just be the start of a slide toward behavioral advertising. Developers of popular podcatchers have access to rich, granular data about listener behavior, with the ability to track each press of the "play" and "pause" buttons.[22] NPR, which distributes a podcast listening app, devised a system it calls "Remote Audio Data," which would share this kind of information to others in the ecosystem, perhaps to sell better targeted advertising.[23] At least as of 2020, reports suggested that the developers of podcatchers were reluctant to embrace this shift, citing privacy concerns.[24]

The biggest shift toward a web-style behavioral business model is the shift away from RSS to streaming audio services. Spotify, for example, has started to aggressively market its podcast offerings, but average users may not appreciate that Spotify is hosting and streaming these audio episodes rather than distributing downloads through RSS, giving them and their advertisers—but not necessarily individual podcasters—detailed data about user behavior.[25]

Similarly, Reddit has slowly introducing behavioral components to its advertising model. In 2017, Reddit began showing ads destined to a specific subreddit to users who had previously visited that subreddit, while elsewhere on the website.[26] This breached the purely contextual business model, by shifting Reddit's model to one that monetized past user behavior in this small way. Later, it introduced "interest targeting," a system by which Reddit sorts individual users into broad interest categories such as "health" or "gaming" based on their prior use of the site. These seem to be part of a broader suite of behavioral advertising models Reddit is rolling out.[27]

If my hypothesis is right and the non-behavioral business models of podcasting and Reddit have helped them avoid a key driver of toxic group formation, they should resist these efforts to move both toward behavioral marketing advertising. Not only does the shift to surveillance-based advertising harm user privacy, it might also shift incentives away from the community-centric, boundary-demarcated generators of relatively healthy communities these services have been.

Lessons for Traditional Social Media

Consider some changes traditional social media can adopt to make their services more like Reddit or podcasts. First, they can try to strengthen the boundaries between different communities. One technological affordance they all share are hashtags, which sometimes develop community-like status such as happened with #blacktwitter or #metoo. Traditional social media could change its amplification algorithms to encourage the use of hashtags, limiting the reach of posts without hashtags.

Once hashtags become more universally used, these services could change their UI to encourage users to view posts by hashtag, rather than in a disaggregated feed. Going even further, these services might appoint moderators, either users or employees, to watch over some of the most popular hashtags. Like Reddit moderators, these individuals might be empowered to set, announce, and enforce norms for the community.

Rather than merely repurpose hashtags, traditional social media can embrace formal community spaces with legible boundaries.[28] In 2021, Twitter introduced "Communities," which some have characterized as an attempt to be more like Reddit.[29] Facebook has long had "Groups," self-contained spaces on the site overseen by moderators.[30]

Both of these sub-services anoint moderators with powers similar to Reddit moderators.

If traditional social media won't take such steps on their own, regulators can nudge them to do so. In the United States, Congress could amend Section 230 to limit immunity from liability to services that have legible boundaries between moderated communities.

Much more aggressively, we could ban behavioral ads, on the theory that contextual advertising would lead to healthier communities. The E.U.'s recent ruling on Facebook is a step in the right direction.

Notes

1 Steven J. Jackson et al., *The Policy Knot: Re-Integrating Policy, Practice and Design in CSCW Studies of Social Computing*, 17 Proc. ACM Conf. on Comput. Supported Coop. Work & Soc. Comput. 588 (2014), https://dl.acm.org/doi/10.1145/2531602.2531674.

2 To be clear, neither podcasts nor Reddit is free from worrisome group dynamics at times, but both seem, at least to me, to create crowds that are healthier and more accountable than those that assemble in services designed on behavioral advertising.

3 *See* Adrianne Jeffries, *Is Your Favorite Podcast Tracking You?*, The Markup (Oct. 8, 2020), https://themarkup.org/the-breakdown/2020/10/08/podcast-privacy-tracking-listener-data.

4 *See* Jeffries, *supra* note 3.

5 *See Feed Validation Service*, W3C (last visited June 9, 2023), https://validator.w3.org/feed/docs/rss2.html.

6 *See* John Sullivan et al., *How Apple's New Audio Subscriptions Are Upending Podcasting*, Fast Co. (May 15, 2021), https://www.fastcompany.com/90636345/how-apples-new-audio-subscriptions-are-upending-podcasting.

7 *See U.S. Podcast Advertising Revenue Report: FY 2021 Results & 2022–2024 Growth Projections*, Interactive Advert. Bureau (May 9, 2022), https://www.iab.com/insights/u-s-podcast-advertising-revenue-report-fy-2021-results-2022-2024-growth-projections/.

8 For curious readers, in early 2023, my favorite subreddits are r/cyberdeck, r/SBCgaming, r/EDC, and r/Emacs.

9 *See* David Curry, *Reddit Revenue and Usage Statistics (2023)*, Bus. Apps (Jan. 9, 2023), https://www.businessofapps.com/data/reddit-statistics/.

10 *See* James Vincent, *Reddit Is Now Valued at More than $10 Billion*, The Verge (Aug. 12, 2021), https://www.theverge.com/2021/8/12/22621445/reddit-valuation-revenue-funding-round.

11 *See* Arvind Narayanan, *TikTok's Secret Sauce*, Knight First Amend. Inst. Colum. U. (Dec. 15, 2022), http://knightcolumbia.org/blog/tiktoks-secret-sauce.

12 *See* Vincent, *supra* note 10.

13 *See* Andrea Peterson, *Reddit Is Finally Cracking Down on Revenge Porn*, Wash. Post (Feb. 24, 2015), https://www.washingtonpost.com/news/the-switch/wp/2015/02/24/reddit-is-finally-cracking-down-on-revenge-porn/.

14 *See* Mike Isaac, *Reddit, Acting Against Hate Speech, Bans "The_Donald" Subreddit*, N.Y. Times (June 29, 2020), https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html.

15 *See, e.g.,* Stephen Marche, *Swallowing the Red Pill: A Journey to the Heart of Modern Misogyny*, The Guardian (Apr. 14, 2016), https://www.theguardian.com/technology/2016/apr/14/the-red-pill-reddit-modern-misogyny-manosphere-men.

16 *See* Daniel Engber, *Joe Rogan's Show May Be Dumb. But Is It Actually Deadly?*, The Atlantic (Feb. 10, 2022), https://www.theatlantic.com/health/archive/2022/02/joe-rogan-covid-vaccine-misinformation/622040/.

17 *See* Aja Romano, *How Do You Solve a Problem like Joe Rogan?*, Vox (Feb. 23, 2022), https://www.vox.com/culture/22945864/joe-rogan-politics-spotify-controversy.

18 Georgina Smith, *Logan Paul Hits Back at His "Toxic & Volatile" Podcast Audience*, Dexerto (Sept. 25, 2020), https://www.dexerto.com/entertainment/logan-paul-hits-back-at-his-toxic-volatile-podcast-audience-1424103/.

19 *See* Myles Allan, *The Rise of Sexism in Indie Podcasts*, Study Breaks (May 23, 2022), https://studybreaks.com/tvfilm/the-rise-of-sexism-in-indie-podcasts/. One participant at the Barons and Mob workshop observed that podcasts had become a way to slander other people without consequence.

20 *See* Jeffries, *supra* note 3.

21 *Acast Leads Global Podcast Advertising Charge by Launching Acast Marketplace*, Acast (Feb. 26, 2020), https://medium.com/acast/acast-leads-global-podcast-advertising-charge-by-launching-acast-marketplace-746ad32c332d.

22 *See Acast Leads Global Podcast Advertising Charge by Launching Acast Marketplace*, *supra* note 21.

23 *See About Remote Audio Data*, NPR (last visited June 9, 2023), https://rad.npr.org/dotorg/about-rad/.

24 *See* Jeffries, *supra* note 3.

25 *See* Jeffries, *supra* note 3 ("Spotify knows if your phone is in your hand or in your pocket, for example.").

26 *See* bold_panda, *Reddit's Ad Changes Reduce Your ROI*, Hacker News (Sept. 12, 2017), https://news.ycombinator.com/item?id=15233127.

27 Two others are called "custom audience," *see Reddit Engagement Retargeting*, Reddit Ads (last visited June 9, 2023), https://redditinc.force.com/helpcenter/s/article/Overview-Custom-Audience-Reddit-

Engagement-Retargeting, and "pixel retargeting," *see Pixel Retargeting*, Reddit Ads (last visited June 9, 2023), https://redditinc.force.com/helpcenter/s/article/Reddit-Ads-Pixel-Retargeting.

28 *See* David Regan, *Communities: Talk About Your Thing with People Who Get You*, Twitter Blog (Sept. 9, 2021), https://blog.twitter.com/en_us/topics/product/2021/testing-communities.

29 *See* Anita George, *Twitter Communities: A Reddit-like World Within Twitter*, Digit. Trends (June 14, 2022), https://www.digitaltrends.com/social-media/twitter-communities-a-reddit-like-world-within-twitter/.

30 *See Groups*, Facebook Help Ctr. (last visited June 9, 2023), https://www.facebook.com/help/1629740080681586.

Acknowledgments

The Barons and the Mob is the product of a virtual workshop on platforms and crowds, held online by the Cornell Tech Research Lab in Applied Law and Technology (CTRL-ALT) on July 28 and August 4, 2022. In addition to the scholars who have authored chapters of this report, participants in the workshop included experts from civil society, media, and industry.

The editors would like to thank all of the participants in the workshop for their contributions to this project. We would also like to thank the students of the Spring 2022 course Directed Reading in Content Moderation at Cornell Tech, who helped to develop content for the workshop including the annotated bibliography of this report. Finally, we would like to thank Microsoft for generously providing financial support for this project.

Views and opinions expressed in this report are those of the individual authors, and should not be attributed to other authors, workshop participants, or affiliated institutions.

About the Authors

Jessica Beyer is an Assistant Teaching Professor in the Henry M. Jackson School of International Studies at the University of Washington. She holds a Ph.D. from the University of Washington, an M.A. in political science from the University of Washington, an M.A. in international studies from the University of Bath, and a B.A. from the University of Washington.

Finn Brunton is a Professor of Science and Technology Studies and of Cinema and Digital Media at the University of California, Davis. He holds a Ph.D. in modern thought from the University of Aberdeen, an M.A. in communications theory from the European Graduate School, and a B.A. from the University of California, Berkeley.

Gabriella Coleman is a Professor of Anthropology at Harvard University. She holds a Ph.D. and M.A. in socio-cultural anthropology from the University of Chicago, and a B.A. from Columbia University.

Evelyn Douek is an Assistant Professor of Law at Stanford Law School. She holds an S.J.D. and LL.M. from Harvard Law School, and a B.Comm. and LL.B. from the University of New South Wales.

Charles Duan is an Assistant Professor of Law at the American University Washington College of Law. He holds a J.D. from Harvard Law School and an A.B. from Harvard College.

James Grimmelmann is the Tessler Family Professor of Digital and Information Law at Cornell Tech and Cornell Law School. He holds a J.D. from Yale Law School and an A.B. from Harvard College.

Nikolas Guggenberger is an Assistant Professor of Law at the University of Houston Law Center. He holds an LL.M. from Stanford Law School and a D. jur. from the University of Freiburg School of Law.

Bing He is a Ph.D. student in the School of Computer Science at the Georgia Institute of Technology. He holds an M.Sc. in computer science from the University of Macau, and a B.E. from the University of Electronic Science and Technology of China.

Srijan Kumar is an Assistant Professor in the School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology. He holds a Ph.D. in computer science from the University of Maryland and a B.Tech. from the Indian Institute of Technology, Kharagpur.

Alice Marwick is an Associate Professor in the Department of Communication at the University of North Carolina at Chapel Hill. She holds a Ph.D. in media, culture, and communication from New York University, an M.A. in communication from the University of Washington, and a B.A. from Wellesley College.

Paul Ohm is a Professor of Law at the Georgetown University Law Center. He holds a J.D. from the University of California, Los Angeles, and a B.S. and B.A. from Yale University.

Rebecca Tushnet is the Frank Stanton Professor of the First Amendment at Harvard Law School. She holds a J.D. from Yale Law School and an A.B. from Harvard College.

Bibliography

Problematic Content

Hate Speech and Incitement

David Kaye, Speech Police: The Global Struggle to Govern the Internet (2019), https://globalreports.columbia.edu/books/speech-police/.

> Argues that the combination of dominant platforms and government regulation of content, primarily outside the United States, are diminishing individual and democratic rights in favor of corporate power.

Margot Kaminski, *Incitement to Riot in the Age of Flash Mobs*, 81 U. Cin. L. Rev. 1 (2013), https://scholarship.law.uc.edu/uclr/vol81/iss1/1.

> Considers application of First Amendment jurisprudence on incitement-to-riot statutes, in view of flash mobs organized on social media.

Harassment

Sarah Banet-Weiser & Kate M. Miltner, *#MasculinitySoFragile: Culture, Structure, and Networked Misogyny*, 16 Feminist Media Stud. 171 (2016), https://sites.middlebury.edu/soan191/files/2018/08/MasculnitysoFragile.pdf.

> Argues that networked misogyny is in part an outgrowth of early expectations of predominantly white, male users that online social spaces were exclusively theirs, giving rise to pushback as women and minorities entered.

Danielle Keats Citron, Hate Crimes in Cyberspace (2014), https://www.hup.harvard.edu/catalog.php?isbn=9780674659902.

In view of widespread and troubling examples of online harassment and attacks particularly against women, proposes reforms toward cyber civil rights that coordinate platforms, civil law, and law enforcement to address such harms.

Kishonna L. Gray, Bertan Buyukozturk & Zachary G. Hill, *Blurring the Boundaries: Using Gamergate to Examine "Real" and Symbolic Violence against Women in Contemporary Gaming Culture, in* 11 Socio. Compass No. e12458 (2017).

Examines the dynamics of Gamergate and places it into the context of marginalization of women in video games as well as offline.

Sarah Jeong, Internet of Garbage (2018), https://www.theverge.com/2018/8/28/17777330/internet-of-garbage-book-sarah-jeong-online-harassment.

Analyzes the mechanisms by which online harassment operates and proliferates, how it might be defined and understood, and why the structure of the internet and the policies related to it are insufficient for addressing it.

Shagun Jhaver et al., *The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action, in* 23 First Monday (2018), https://firstmonday.org/ojs/index.php/fm/article/view/8232/6644.

Interviews posters on a Reddit forum actively involved in GamerGate, to assess their norms and beliefs about harassment.

Amanda Lenhart et al., Online Harassment, Digital Abuse, and Cyberstalking in America (Data & Soc'y Nov. 21, 2016), https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/.

Based on survey data, finds extensive evidence of online harassment and self-censorship that results.

Rebecca Lewis, Alice E. Marwick & William Clyde Partin, *"We Dissect Stupidity and Respond to It": Response Videos and Networked Ha-*

rassment on YouTube, 65 Am. Behav. Scientist 735 (2021), https://journals.sagepub.com/doi/abs/10.1177/0002764221989781.

Documents techniques by which online harassers and their communities of support take advantage of the affordances of YouTube to avoid removal.

Alice E. Marwick, *Morally Motivated Networked Harassment as Normative Reinforcement, in* 7 Soc. Media & Soc'y (2021), https://journals.sagepub.com/doi/full/10.1177/20563051211021378.

Provides a model of networked harassment as morally motivated, prompted through creation of moral justifications for harassment that highlight competing moral claims and norms between communities.

Alice E. Marwick & Robyn Caplan, *Drinking Male Tears: Language, the Manosphere, and Networked Harassment*, 18 Feminist Media Stud. 543 (2018), http://www.tiara.org/wp-content/uploads/2018/05/Marwick_Caplan_Drinking-male-tears-language-the-manosphere-and-networked-harassment.pdf.

Examines the self-reinforcing network of men's rights advocates known as the "manosphere," and the effect of that network on perpetuating anti-feminist harassment.

Adrienne Massanari, *#Gamergate and the Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures*, 19 New Media & Soc'y 329 (2015), https://journals.sagepub.com/doi/abs/10.1177/1461444815608807.

Reviews how Reditt's technological affordances allowed it to become a hub of anti-feminist and misogynistic activism.

Sarah Sobieraj, Credible Threat: Attacks Against Women Online and the Future of Democracy (2020).

Discusses the digital harassment women face when they are vocal about political and social issues—both the types and severity as well as the broader democratic consequences.

Nonconsensual Pornography

Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 Wake Forest L. Rev. 345 (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2368946.

Proposes criminalizing revenge porn, namely nonconsensual publication of private graphic images, and argues that such criminalization fits within the First Amendment.

Clare McGlynn et al., *Beyond "Revenge Porn": The Continuum of Image-Based Sexual Abuse*, 25 Feminist Legal Stud. 25 (2017), https://link.springer.com/article/10.1007/s10691-017-9343-2.

Expands upon the dialogue on revenge porn to identify a range of "image-based sexual abuse" that legislation and other policy measures ought to deal with.

Ari Ezra Waldman, *Law, Privacy, and Online Dating: "Revenge Porn" in Gay Online Communities*, 44 Law & Soc. Inquiry 987 (2019), https://www.cambridge.org/core/journals/law-and-social-inquiry/article/law-privacy-and-online-dating-revenge-porn-in-gay-online-communities/BCCE05CF25AA4C2E05CCF8D64980E839.

Traces the high prevalence of revenge porn in such communities to platform norms that all but require disclosure of explicit photos and to a failure of self-help techniques to maintain privacy in light of those norms.

Misinformation and Political Manipulation

Mia Bloom & Sophia Moskalenko, Pastels and Pedophiles: Inside the Mind of QAnon (2021), http://www.sup.org/books/title/?id=34673.

Describes the development and workings of the ongoing conspiracy theory narrative and its followers, with particular attention to anti-Semitic origins and associations with gender.

R. Kelly Garrett & Shannon Poulsen, *Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warn-*

ings, 24 J. Comput.-Mediated Commc'n 240 (Sept. 1, 2019), https://academic.oup.com/jcmc/article/24/5/240/5575583.

Compares different mechanisms of flagging content for their effectiveness at reducing sharing and belief.

Matthew Hannah, *QAnon and the Information Dark Age*, *in* 26 First Monday No. 2 (Feb. 1, 2021), https://journals.uic.edu/ojs/index.php/fm/article/view/10868.

Argues that expanded communication technologies, in combination with skepticism of academia and mainstream media, have enabled the conspiratorial group thinking that characterizes QAnon.

Sumeet Kumar et al., *An Anatomical Comparison of Fake-News and Trusted-News Sharing Pattern on Twitter*, 27 Comput. & Mathematical Org. Theory 109 (June 1, 2021), https://link.springer.com/article/10.1007/s10588-019-09305-5.

Finds trends in hashtag uses, mentions, and negative sentiments in Twitter sharing of misinformation in Ukraine.

Anthony Nadler et al., Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech (Data & Soc'y 2009), https://www.datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf.

Evaluates the role and impact of targeted digital advertising infrastructures, arguing that their outsized influence enables problematic political manipulation.

Whitney Phillips, The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online (Data & Soc'y 2018), https://datasociety.net/library/oxygen-of-amplification/.

In the context of the 2016 presidential election, analyzes how media amplification lent visibility and legitimacy to far-right fringe groups that exploited "classic" trolling strategies of manipulation.

Elizabeth Stewart, *Detecting Fake News: Two Problems for Content Moderation*, 34 Phil. & Tech. 923 (2021), https://pubmed.ncbi.nlm.nih.gov/33589871/.

Observes that moderation to limit fake news requires value judgments that can generate user distrust toward fact-checking efforts, an unavoidable problem that could be mitigated with diverse expert moderators and increased transparency.

Daniel Susser et al., *Online Manipulation: Hidden Influences in a Digital World*, 4 Geo. L. Tech. Rev. 1 (2019), https://georgetownlawtechreview.org/wp-content/uploads/2020/01/4.1-p1-45-Susser.pdf.

Explores the connection between "manipulation," defined as the act of commandeering others' decisionmaking processes, and the capabilities of digital technologies.

Francesca Tripodi, Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices (Data & Soc'y May 16, 2018), https://datasociety.net/library/searching-for-alternative-facts/.

Based on ethnographic study of two Republican groups, draws connections between conservative Christian scriptural inference practices and approaches to news media and fact-checking.

Renée DiResta, Invisible Rulers: The People Who Turn Lies Into Reality (2024).

Describes coordinated networks of propagandists who shape public opinion and undermine trust in institutions.

Deep Fakes

Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Calif. L. Rev. 1753 (2019), http://www.californialawreview.org/wp-content/uploads/2020/01/2-Chesney-Citron.34.final_.pdf.

Identifies potential concerns with deep fake video technology, in which artificial intelligence systems generate realistic content, and reviews policy responses to such concerns.

Britt Paris & Joan Donovan, Deepfakes and Cheap Fakes (Data & Soc'y Sept. 18, 2019), https://datasociety.net/library/deepfakes-and-cheap-fakes/.

> Questions concerns with deep fake videos by contextualizing them within other misleading audiovisual content.

Terrorism

Bennett Clifford, Moderating Extremism: The State of Online Terrorist Content Removal Policy in the United States (Program on Extremism, George Wash. Univ. 2021), https://extremism.gwu.edu / sites / g / files / zaxdzs2191 / f / Moderating %20Extremism %20The %20State %20of %20Online %20Terrorist %20Content %20Removal%20Policy%20in%20the%20United%20States.pdf.

> Questions effectiveness of social media regulation for combatting terrorist and extremist content.

Stuart MacDonald et al., *Regulating Terrorist Content on Social Media: Automation and the Rule of Law*, 15 Int'l J.L. Context 183 (2019), https: / / www.cambridge.org / core / journals / international - journal - of - law - in - context / article / regulating - terrorist - content - on - social - media - automation - and - the - rule - of - law/B54E339425753A66FECD1F592B9783A1.

> Argues that automated detection and removal of terrorist content raises issues including displacement (moving to other platforms upon being banned), defining terrorism, and appeal rights.

Alice Marwick, Benjamin Clancy & Katherine Furl, Far-Right Online Radicalization: A Review of the Literature (Ctr. for Info., Tech., & Pub. Life 2022), https://assets.pubpub.org/9694oeej/31648241763479.pdf.

> Based on a review of literature on radicalization theories primarily developed in the post-9/11 context, argues that those theories are not well-suited to conceptualizing growth of the fring far-right.

Intellectual Property

Casey Fiesler, *Everything I Need to Know I Learned from Fandom: How Existing Social Norms Can Help Shape the Next Generation of User-Generated Content*, 10 Vand. J. Ent. & Tech. L. 729 (2008), https://papers.ssrn.com/abstract=3634582.

> Argues that copyright law should adapt to fan-created works in ways consistent with existing norms of the community of existing writers of such works.

Casey Fiesler & Amy S. Bruckman, *Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement*, *in* 3 Proc. ACM on Hum.–Comput. Interaction No. 241 (2019), https://cmci.colorado.edu/~cafi5706/group2020_fiesler.pdf.

> Identifies social norms of copying and remixing among fandom communities, and the effectiveness of such rules.

Jennifer M. Urban et al., Notice and Takedown in Everyday Practice (ver. 2 Mar. 2017), https://papers.ssrn.com/abstract=2755628.

> Studies of notice and takedown regime of the Digital Millennium Copyright Act finding widespread recognition of its role in the online ecosystem but also frequent human and automated errors in decisionmaking.

Spam

Finn Brunton, Spam: A Shadow History of the Internet (2013), https://mitpress.mit.edu/books/spam.

> By tracing the history of spam and anti-spam practices, shows how email, search engines, and online communities' governance models have been shaped by adapting to the problem.

Content Moderation

How Moderation Happens

Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches (Data & Soc'y

Nov. 14, 2018), https://datasociety.net/library/content-or-context-moderation/.

Taxonomizes content moderation approaches that platforms adopt, and identifies tradeoffs between consistency and localized context sensitivity in these approaches.

Jialun Aaron Jiang et al., *Moderation Challenges in Voice-based Online Communities on Discord*, *in* 3 Proc. ACM on Hum.–Comput. Interaction No. 55 (2019).

Discusses strategies and challenges for moderating content on real-time voice-based communication systems, particularly with respect to hate speech and harassment.

Tarleton Gillespie, Custodians of the Internet, Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media (2021), https://yalebooks.yale.edu/9780300261431/custodians-of-the-internet.

Argues that content moderation is a "fundamental part" of platforms' functioning and success, which forces a reframing of platforms' role in culture and society.

Eric Goldman, *Content Moderation Remedies*, 28 Mich. Tech. L. Rev. 1 (2021), https://repository.law.umich.edu/mtlr/vol28/iss1/2.

Proposes that social media platforms become less punitive and instead reduce harmful content on their sites by adopting other remedies that give users control over their experience and the content they see.

James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42 (2015), https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2620&context=facpub.

Proposes a taxonomy of techniques, distinctions, and community characteristics to capture a diversity of approaches to content moderation.

Governance and Decisionmaking

Kate Crawford & Tarleton Gillespie, *What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Soc'y 410 (2014), https://papers.ssrn.com/abstract=2476464.

Characterizes flagging as both a practical content moderation tool and a "rhetorical mechanism" that aids in legitimizing a site's content.

Evelyn Douek, *Content Moderation as Systems Thinking*, 136 Harv. L. Rev. 526 (2022).

In view of the complexity and dynamicness of modern content moderation, argues that regulatory policy must focus less on substantive rules and decisions, and more on administrative infrastructures and institutional designs surrounding moderation.

Robert Gorwa, *What Is Platform Governance?*, 22 Info. Commc'n & Soc'y 854 (2019), https://gorwa.co.uk/files/platformgovernance.pdf.

Taxonomizes models of governance for online platforms into self-governance, external governance directed by legislation or regulation, and co-governance through public participation or partnerships.

Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2018), https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

Examines the policies and procedures of content moderation on major platforms, as well as the influences that shape them, to argue that platform content moderation is best understood as a system of governance.

Facebook Oversight Board

Evelyn Douek, *Facebook's "Oversight Board": Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & Tech. 1 (2019), https://scholarship.law.unc.edu/ncjolt/vol21/iss1/2.

Proposes that the primary objective and value of the Facebook Oversight Board should be to highlight weaknesses with Facebook's moderation policies and to enable a process of "public reasoning" that legitimizes the decisionmaking process.

Evelyn Douek, *The Facebook Oversight Board's First Decisions: Ambitious, and Perhaps Impractical*, Lawfare (Jan. 28, 2021), https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical.

> Reviews the Facebook Oversight Board's first five decisions as opening a dialogue with the company not just on individual moderation decisions but on broader policies and procedures.

Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L.J. 2418 (2020), https://www.yalelawjournal.org/feature/the-facebook-oversight-board.

> Characterizes the history of the Facebook Oversight Board and identifies the Board's potential utility in engaging user participation in platform governance.

Algorithmic Moderation

Michael A. DeVito et al., *"Algorithms Ruin Everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media*, 2017 Proc. CHI Conf. on Hum. Factors Comput. Sys. 3163, https://michaelanndevito.files.wordpress.com/2017/01/riptwitter_chi2017.pdf.

> Investigates reasons for public resistance to algorithmic content curation systems such as news feeds.

Tarleton Gillespie, *Content Moderation, AI, and the Question of Scale*, *in* 2020 Big Data & Soc'y, https://journals.sagepub.com/doi/full/10.1177/2053951720943234.

> Questions use of artificial intelligence in content moderation and argues that automated moderation technologies ought to be used as an adjunct to human decisionmaking rather than a replacement.

Zeynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 Colo. Tech. L.J.

203 (2015), https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf.

Argues that the social, emotional, and political consequences of content moderation demand greater scrutiny into the algorithms that perform such moderation.

Jing Zeng & D. Bondy Valdovinos Kaye, *From Content Moderation to Visibility Moderation: A Case Study of Platform Governance on TikTok*, 2022 Pol'y & Internet 79, https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.287.

Discusses TikTok's use of "visibility moderation," in which content is moderated through algorithmic adjustment of its public visibility.

Intermediary Liability

Ctr. for Democracy & Tech., Shielding the Messengers: Protecting Platforms for Expression and Innovation (Dec. 2012), https://cdt.org/insights/shielding-the-messengers-protecting-platforms-for-expression-and-innovation/.

Argues that intermediary immunity from liability for user-generated content expands online expression, encourages innovation, and creates more opportunities for local content, thereby supporting development of the information society.

The Oxford Handbook of Online Intermediary Liability (Giancarlo Frosio ed., 2020), https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780198837138.001.0001/oxfordhb-9780198837138.

Overview of intermediary liability issues with particular focus on international differences, theories of liability, and current policy trends.

Eric Goldman, *The Ten Most Important Section 230 Rulings*, 20 Tul. J. Tech. & Intell. Prop. 1 (2017), https://journals.tulane.edu/TIP/article/view/2676.

Discusses the impact, importance, and controversy of section 230 through prominent case law.

James Grimmelmann, *To Err Is Platform: Response to Olivier Sylvain's Essay "Discriminatory Designs on User Data"* (Apr. 6, 2018), https://knightcolumbia.org/content/err-platform.

> Identifies a key difficulty with section 230 reform proposals as indeterminacy about how to separate good and bad content accurately at scale.

Daphne Keller, Hoover Institution, Aegis Series Paper No. 1807, Internet Platforms: Observations on Speech, Danger, and Money (2018), https://www.hoover.org/research/internet-platforms-observations-speech-danger-and-money.

> Considers consequences of over-removal incentives of platforms that could arise from limiting intermediary immunity, consequences that include harms to free expression, security, and innovation.

Jeff Kosseff, The Twenty-Six Words That Created the Internet (2019), https://www.cornellpress.cornell.edu/book/9781501714412/the-twenty-six-words-that-created-the-internet/.

> Reviews the history and impact of section 230.

Note, *Section 230 as First Amendment Rule*, 131 Harv. L. Rev. 2027 (2018), https://harvardlawreview.org/2018/05/section-230-as-first-amendment-rule/.

> Explains courts' generally broad interpretation of section 230 immunity as a consequence of the law's close tie to the First Amendment.

Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 Notre Dame L. Rev. 293 (2011), https://scholarship.law.nd.edu/cgi/viewcontent.cgi?article=1005&context=ndlr.

> Identifies the problem of platforms over-suppressing lawful speech in order to ensure compliance with filtering mandates, a problem that ought to inform interpretation of section 230.

Cross-Cultural Understanding

Mahsa Alimardani & Mona Elswah, *Digital Orientalism: #SaveSheikh-Jarrah and Arabic Content Moderation*, Project on Middle E. Pol. Sci.

Geo. Wash. U. (Aug. 5, 2021), https://pomeps.org/digital-orientalism-savesheikhjarrah-and-arabic-content-moderation.

Based on analysis of Arabic content moderation policies, argues that current major platforms exhibit "digital orientalism" that stereotypes views of the Middle East and North Africa.

Michael Karanicolas, *Moderate Globally, Impact Locally: A Series on Content Moderation in the Global South*, Yale Info. Soc'y Project (Aug. 5, 2020), https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-series-content-moderation-global-south.

Blog post series on content moderation in countries and regions including Russia, Myanmar, Africa, Latin America, and India, with particular focus on places with limited traditional media.

Jeffrey Sablosky, *Dangerous Organizations: Facebook's Content Moderation Decisions and Ethnic Visibility in Myanmar*, 43 Media Culture & Soc'y 1017 (2021), https://journals.sagepub.com/doi/abs/10.1177/0163443720987751.

Documents Facebook's banning of combatant organizations in Myanmar as exemplary of how the platform's content moderation practices can cut off ethnic groups from national and international visibility.

Tom Simonite, *Facebook Is Everywhere; Its Moderation Is Nowhere Close*, Wired (Oct. 25, 2021), https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/.

Explores issues that Facebook has faced with respect to Arabic content.

Zahra Takhshid, *Regulating Social Media in the Global South*, 24 Vand. J. Ent. & Tech. L. 1 (2021), https://scholarship.law.vanderbilt.edu/jetlaw/vol24/iss1/1.

Argues that U.S. firms' lack of attention to concerns of the Global South requires collective action through regional treaties and other means.

Platform Cooperation

Evelyn Douek, The Rise of Content Cartels (Knight First Amend. Inst. at Columbia Univ. 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels.

> Considers normative consequences of platform cooperation, particularly that which occurs behind the scenes among major platforms, identifying concerns of accountability, dominance, and public trust.

The Manila Principles on Intermediary Liability Background Paper (2015), https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf.

> Multistakeholder-developed principles for content moderation intended to guide government, industry, and civil society.

Government–Platform Interactions

Hannah Bloch-Wehba, *Content Moderation as Surveillance*, 36 Berkeley Tech. L.J. 1297 (2021), https://btlj.org/wp-content/uploads/2023/01/0012-36-3-Bloch-Wehba_Web.pdf.

> Explores the relationship between platforms as content moderators and law enforcement, which seeks to influence moderation decisions but also challenges platforms' power.

Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. Rev. 27 (2019), https://scholar.smu.edu/smulr/vol72/iss1/9/.

> Contends that governments worldwide are coopting platforms' private arrangements in order to implement those governments' public policy preferences, which creates problems of extraterritoriality, accountability, and transparency.

Brian Chang, *From Internet Referral Units to International Agreements; Censorship of the Internet by the UK and EU*, 49 Colum. Hum. Rts. L. Rev. 114 (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3062909.

> Describes U.K. and E.U. Internet Referral Units, government entities that advise platforms to remove terrorist or extremist content, and their implications for speech and human rights law.

Moderation Working Conditions

Casey Newton, *The Secret Lives of Facebook Moderators in America*, The Verge (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

Describes Facebook content moderator workers' difficulties and lack of support from their employer.

Sarah T. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media (2021), https://yalebooks.yale.edu/9780300261479/behind-the-screen.

Documents the working conditions, experiences, and exploitation of those in the content moderation industry.

Dynamics of Crowds

Virality, Community Building, and Trolling

Jessica L. Beyer, *Women's (Dis)embodied Engagement with Male-Dominated Online Communities*, *in* Cyberfeminism 2.0, at 153 (Radhika Gajjala & Yeon Ju Oh eds., 2012).

Explores collective player-created barriers to women's equal participation in World of Warcraft and other online gaming spaces based on stereotypical beliefs about women, facilitated by the anonymity and disembodied nature of the online space as well as content moderation practices.

Jean Burgess, *"All Your Chocolate Rain Are Belong to Us"?: Viral Video, YouTube and the Dynamics of Participatory Culture*, *in* Video Vortex Reader: Responses to YouTube 101 (Geert Lovink & Sabine Niederer eds., 2008).

Examines viral content and the dynamics of participatory culture through the spread of two YouTube videos, both of which went viral through different mechanisms and types of crowd behavior.

E. Gabriella Coleman, *Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle*, *in* The Social Media Reader 44 (Michael

Mandiberg ed., 2012), http://media-study.com/resources/pdfs/socialmedia.pdf.

Traces evolution of online trolling through early computer hacking, arguing that trolls are motivated by antagonism toward "massification of the Internet."

Kristen Engel et al., *Characterizing Reddit Participation of Users Who Engage in the QAnon Conspiracy Theories*, *in* 6 Proc. ACM on Hum.–Comput. Interaction No. 53 (2022), https://dl.acm.org/doi/pdf/10.1145/3512900.

Analyzes usage patterns of Reddit posters active on QAnon subreddits surrounding Reddit's 2018 ban of those subreddits.

Kishonna L. Gray, *Deviant Bodies, Stigmatized Identities, and Racist Acts: Examining the Experiences of African-American Gamers in Xbox Live*, 18 New Rev. Hypermedia & Multimedia 261 (2012).

Illustrates how the broader community of Xbox Live labels and treats minority gamers, particularly African-American males, as deviant through linguistic profiling.

James Grimmelmann, *The Platform Is the Message*, 2 Geo. L. Tech. Rev. 217 (2018), https://georgetownlawtechreview.org/the-platform-is-the-message/GLTR-07-2018/.

The popularity and virality of seemingly undesirable content online is a result of the economic and technical structure of the Internet and social media platforms: the speed, scale, and fidelity with which content can be reproduced; platforms' algorithmic rewarding of extreme content; disconnects between a content poster's intentions and downstream readers' interpretations; and monetization incentives of platforms.

Johannes Loh & Tobias Kretschmer, *Platform Competition and User-Generated Content: Evidence from Game Wikis* (Ctr. for Econ. Pol'y Rsch., Discussion Paper DP16107, May 4, 2020), https://repec.cepr.org/repec/cpr/ceprdp/DP16107.pdf.

For online platforms that depend on communities of unpaid contributors, finds that increased dominance of a platform correlates with greater contributor activity.

Alice E. Marwick & William Clyde Partin, *Constructing Alternative Facts: Populist Expertise and the QAnon Conspiracy*, *in* 2022 26 New

Media & Soc'y 2535, https://journals.sagepub.com/doi/full/10.1177/14614448221090201.

Argues that QAnon participants are not gullible fools but construct and maintain their own episteme through leveraging alternative truth claims, called populist expertise.

Whitney Phillips, This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture (2015).

Argues that online trolling is closely tied to traditional media institutions and cultural norms, such that the former is not easily distinguished from the latter.

Whitney Phillips & Ryan M. Milner, The Ambivalent Internet: Mischief, Oddity and Antagonism Online (2017), https://www.wiley.com / en - us / The+Ambivalent+Internet %3A+Mischief %2C+Oddity%2C+and+Antagonism+Online-p-9781509501274.

Explores Internet content from the lens of folklore, arguing that its ambivalent nature—simultaneously community-building and divisively antagonistic—complicates questions such as free expression and media amplification.

Attention, Influence, and Commerce

Yiqing Hua et al., *Characterizing Alternative Monetization Strategies on YouTube*, *in* 6 Proc. ACM on Hum.-Comput. Interaction art. 283 (2022), https://dl.acm.org/doi/pdf/10.1145/3555174.

Surveys prevalence of YouTube videos' revenue sources separate from platform-provided video advertising, and considers implications for platform gatekeeper control, content quality, and rhetorical polarization.

Alice Marwick & Rebecca Lewis, Media Manipulation and Disinformation Online (Data & Soc'y May 15, 2017), https://datasociety.net/library/media-manipulation-and-disinfo-online/.

Documents techniques of "attention hacking" to spread ideas through both social and traditional media, reasons why such techniques are effective, and their exploitation to popularize harmful content, radicalization, and disinformation.

Alice E. Marwick, Status Update: Celebrity, Publicity, and Branding in the Social Media Age (2013), https://yalebooks.yale.edu/9780300209389/status-update.

> Argues that the design of social media, reflective of the free-market competitiveness of startup founders, elevates personal status in ways that promote inequality.

Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads (2017), https://www.penguinrandomhouse.com/books/234876/the-attention-merchants-by-tim-wu/.

> Argues that the current debate about the attention economy fits into a long-term trend of communications industries' increasing and increasingly effective efforts to capture consumers' attention. Those efforts have become especially potent today, Wu contends, with the granular and detailed information on individual users that online technologies can provide.

Bots and Manufactured Popularity

Paul Charon & Jean-Baptiste Jeangéne Vilmer, Chinese Influence Operations: A Machiavellian Moment (Oct. 2021), https://www.irsem.fr/report.html.

> Chapter IX describes China's use of techniques such as sock-puppet accounts, bots, and trolling to spread disinformation to both Chinese citizens and the rest of the world.

Nicholas Confessore et al., *The Follower Factory*, N.Y. Times, Jan. 27, 2018, https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

> Investigates the "shadowy global marketplace" for fake-account followers on Twitter and other platforms, which often involves fraud and identity theft.

Deen Freelon et al., *Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation*, 40 Soc. Sci. Comput.

Rev. 560 (2020), https://journals.sagepub.com/doi/abs/10.1177/0894439320914853.

Finds that racial misrepresentation, particularly Russian election disruption campaigns posing as Black activists, strongly predicts disinformation effectiveness.

Sanjana Hattotuwa et al., Weaponising 280 Characters: What 200,000 Tweets and 4,000 Bots Tell Us About State of Twitter in Sri Lanka (Apr. 21, 2018), https://www.cpalanka.org/wp-content/uploads/2018/04/Weaponising-280-characters.pdf.

Describes activities and influence of bots with respect to the 2018 riots in Sri Lanka.

Ross A. Malaga, *Worst Practices in Search Engine Optimization*, 51 Commc'ns ACM 147 (2008), https://dl.acm.org/doi/fullHtml/10.1145/1409360.1409388.

Discusses problematic techniques for manipulating search engine rankings.

Stanford Internet Observatory, Reply-Guys Go Hunting: An Investigation into a U.S. Astroturfing Operation on Facebook, Twitter, and Instagram (Oct. 8, 2020), https://stacks.stanford.edu/file/druid:vh222ch4142/facebook-US-202009.pdf.

Analyzes inauthentic accounts created by a U.S. media consultancy as part of advocacy campaigns, to identify techniques of coordinated inauthentic behavior.

Crowd-Based Production

Brett M. Frischmann, Infrastructure: The Social Value of Shared Resources (2012), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2000962.

Analyzes demand for infrastructure as a commons resource, calling for open and nondiscriminatory sharing of infrastructure including in fields such as intellectual property and Internet policy.

Yochai Benkler, The Wealth of Networks: How Social Production Transforms Markets and Freedom (Yale Univ. Press 2006), http://www.benkler.org/Benkler_Wealth_Of_Networks.pdf.

Considers implications of the shift to information goods, particularly those produced by individuals and large-scale cooperative efforts, on individuals and society.

Yochai Benkler, *A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate*, 46 Harv. C.R.-C.L.L. Rev. 311 (2011), https://dash.harvard.edu/handle/1/10900863.

Characterizes government and industry responses to Wikileaks as indicative of a new "extralegal public-private partnership" developing to respond to new networked media.

Amy S. Bruckman, Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge (2022), https://www.cambridge.org/core/books/should-you-believe-wikipedia/F1797AA6843FEB206C2D7E418553C39C.

Explores Wikipedia as a case study in online collective action for knowledge development, particularly in comparison to traditional models of research and publication.

Dariusz Jemielniak, Common Knowledge? An Ethnography of Wikipedia (2014), https://www.sup.org/books/title/?id=24010.

Describes Wikipedia's unique model of content moderation and governance arising out of its user-driven bureaucratic structure.

Community Self-Governance

Johan Farkas & Christina Neumayer, *"Stop Fake Hate Profiles on Facebook": Challenges for Crowdsourced Activism on Social Media*, *in* 22 First Monday No. 9 (2017), https://firstmonday.org/ojs/index.php/fm/article/view/8042.

Examines Danish Facebook users' organized efforts to combat ethnic hatred and fake profiles, and considers limits of the sustainability of such crowdsourced efforts.

R. Stuart Geiger, *Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space*, 19 Info.

Commc'n & Soc'y 787 (2016), https://papers.ssrn.com/abstract=2761503.

Proposes use of third-party bot-based collective blocklists as a strategy for content moderation and particularly for increased user control.

Sarita Schoenebeck & Lindsay Blackwell, *Reimagining Social Media Governance: Harm, Accountability, and Repair*, 23 Yale J.L. & Tech. 113 (2021), https://www.justicehappenshere.yale.edu/s/Reimagining-Social-Media-Governance_Harm-Accountability-and-Repair.pdf.

Based on theories of restorative and transformative justice, proposes community-based governance of content moderation as preferable to top-down blocking of offenders.

Hue Watson et al., *"We Hold Each Other Accountable": Unpacking How Social Groups Approach Cybersecurity and Privacy Together*, *in* 2020 Proc. CHI Conf. on Hum. Factors Comput. Sys., https://sauvikdas.com/uploads/paper/pdf/23/file.pdf.

Identifies, based on a survey of small social groups within larger platforms, implicit group norms for protecting privacy and cybersecurity arising from group members' expectations of accountability, responsibility, and trust.

Activism

Jessica L. Beyer, Expect Us: Online Communities and Political Mobilization (2014), https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199330751.001.0001/acprof-9780199330751.

Argues that structures of online communities, in particular anonymity, lack of regulation, and minimal small-group interaction, foster coordinated political mobilization from those communities.

Jessica L. Beyer, *Trolls and Hacktivists: Political Mobilization from Online Communities*, *in* The Oxford Handbook of Digital Media Sociology 417 (Deanna A. Rohlinger & Sarah Sobieraj eds., 2022).

Examines online communities as sites of political mobilization and the interdisciplinary literature that examines them, with focus on four cases

of mobilization: 4chan and trolling culture, Anonymous, Gamergate, and the 2016 U.S. presidental election.

Jessica L. Beyer & Fenwick McKelvey, *You Are Not Welcome Among Us: Pirates and the State*, 9 Int'l J. Commc'n 890 (2015).

Examines political mobilization around digital piracy through a historical review of peer-to-peer networking innovations, in view of contemporary ideals of collaborative culture, nonhierarchical organization, and a reliance on the network.

Jessie Daniels, Cyber Racism: White Supremacy Online and the New Attack on Civil Rights (2009).

Explores how white supremacist organizations have moved online, how they use internet affordances to recruit, and their effectiveness.

Pierce Alexander Dignam & Deana A. Rohlinger, *Misogynistic Men Online: How the Red Pill Helped Elect Trump*, 44 Signs: J. Women Culture & Soc'y 589 (2019).

Analyzes the transformation of the "Red Pill" online forum from an apolitical forum to a key 2016 election influence, as forum leaders reframed a neoliberal and misogynist collective identity to Trump's political platform.

Jennifer Earl & Katrina Kimport, Digitally Enabled Social Change: Activism in the Internet Age (2011).

Articulates differences in online and offline political activity, with focus on two key affordances of reduced costs for protest and decreased need to be physically together to collectively act.

Deen Freelon, Charlton D. McIlwain & Meredith Clark, Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice (Ctr. for Media & Soc. Impact 2016), https://cmsimpact.org/resource/beyond-hashtags-ferguson-blacklivesmatter-online-struggle-offline-justice/.

Discusses the ways in which Black Lives Matter used online tools to organize, mobilize, and press for political change.

Eric Gordon et al., *Why We Engage: How Theories of Human Behavior Contribute to Our Understanding of Civic Engagement in a Digital*

Era (Berkman Ctr. for Internet & Soc'y at Harvard Univ., Research Pub. No. 2013-21, Oct. 2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2343762.

Considers effect of technology on three factors that motivate civic engagement: community trust, senses of political empowerment, and knowledge of engagement opportunities.

J. Nathan Matias, *Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout*, 2016 Proc. CHI Conf. on Hum. Factors Comput. Sys. 1138, https://natematias.com/media/GoingDark-Matias-2016.pdf.

Documents Reddit moderators' 2015 protest of the company's lack of communication with them, to assess the mechanisms of collective action and mobilization that enabled the protest to grow and move rapidly.

An Xiao Mina, Memes to Movements: How the World's Most Viral Media Is Changing Social Protest and Power (2019), https://www.penguinrandomhouse.com/books/567159/memes-to-movements-by-an-xiao-mina/.

Describes how Internet memes, because of their social and communicative power, have become a central tool in political discourse and social movements.

Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest (2018), https://yalebooks.yale.edu/9780300234176/twitter-and-tear-gas.

Investigates the role and limitations of Internet technologies in modern protests and social movements.

Sarah Myers West, *Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms*, *in* 5 Media & Commc'n No. 3, at 28 (2017), https://www.cogitatiopress.com/mediaandcommunication/article/view/989.

Examines a campaign to change Facebook's nudity policy as an example of collective action taking advantage of the affordances of social media platforms to influence the platforms themselves.

Sulafa Zidani, *Represented Dreams: Subversive Expressions in Chinese Social Media as Alternative Symbolic Infrastructures*, *in* 4 Soc. Media

+ Soc'y No. 4 (2018), https://journals.sagepub.com/doi/full/10.1177/2056305118809512.

Investigates use of "subversive expressions" of Chinese Internet users to circumvent state censorship, asking whether such expressions construct a community capable of calling for political action.

Cross-Platform Activity

Zeve Sanderson et al., *Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform*, *in* 2 Harv. Kennedy Sch. Misinformation Rev. No. 4 (2021), https://misinforeview.hks.harvard.edu/wp-content/uploads/2021/08/sanderson_twitter_trump_election_20210824.pdf.

Discusses limited efficacy of Twitter flags and warning labels on misinformation tweets that were shared on platforms outside of Twitter.

Anonymity and Identity Disclosure

David Auerbach, *Anonymity as Culture: Treatise*, Triple Canopy, Feb. 9, 2012, https://www.canopycanopycanopy.com/issues/15/contents/anonymity_as_culture__treatise.

Connects key features of early Internet message boards, in particular anonymity and transience of messages, with the cultures of elitism and trolling that developed in those forums.

Gabriella Coleman, Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous (2015), https://www.versobooks.com/books/2027-hacker-hoaxer-whistleblower-spy.

Characterizes the motivations, development, group dynamics, and actions of the "hacktivist collective."

James Grimmelmann, *Saving Facebook*, 94 Iowa L. Rev. 1137 (2009), https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=2415&context=fac_pubs.

In view of tensions between privacy and the ways in which people use social media sites, identifies and critiques different approaches to social media platform privacy.

Oliver L. Haimson & Anna Lauren Hoffmann, *Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities*, *in* 21 First Monday No. 6 (June 2016), https://journals.uic.edu/ojs/index.php/fm/article/view/6791.

Criticizes Facebook's "authentic identities" policy as paradoxical in that for individuals with marginalized, fluid, or non-normative identities, the policy restricts them from presenting authentic identities.

Christina A. Madsen et al., *Tensions in Scaling-up Community Social Media: A Multi-Neighborhood Study of Nextdoor*, 2014 Proc. SIGCHI Conf. on Hum. Factors Comput. Sys. 2329, https://www.cc.gatech.edu/home/keith/pubs/nextdoor-chi2014.pdf.

Based on a survey of users of a local-neighborhood social media platform, considers how inclusion of users' verified physical locations in social media profiles affects perceptions of privacy and platform design choices.

Helen Nissenbaum, *The Meaning of Anonymity in an Information Age*, 15 Info. & Soc'y 141 (1994), https://nissenbaum.tech.cornell.edu / papers / The %20Meaning %20of %20Anonymity %20in %20an %20Information%20Age.pdf.

In view of the capabilities of information technologies, calls for redefining anonymity beyond mere withholding of names, to the ability of an individual to act while remaining unreachable.

Legal Reforms and Policy Debates

Dominance and Competition in General

David S. Evans et al., Invisible Engines: How Software Platforms Drive Innovation and Transform Industries (2006), https://mitpress.mit.edu/books/invisible-engines.

Contends that the critical feature of the modern technology industry is the two-sided nature of software platforms, which enabled new ways for firms to exploit network effects.

Ben Thompson, *A Framework for Regulating Competition on the Internet*, Stratechery (Dec. 9, 2019), https://stratechery.com/2019/a-framework-for-regulating-competition-on-the-internet/.

> Argues for a regulatory distinction between platforms, which are essential for third party developers and more suited to data regulations, and aggregators that facilitate content discovery, for which mergers and acquisitions should be the focus of regulation.

Tim Wu, The Master Switch: The Rise and Fall of Information Empires (2011), https://scholarship.law.columbia.edu/books/176/.

> Traces the history of development of telecommunications services to show a common "cycle" from open competition to closed dominant firms, a cycle into which digital platforms could potentially fall.

Antitrust Law

David S. Evans, *Governing Bad Behavior by Users of Multi-Sided Platforms*, 27 Berkeley Tech. L.J. 1201 (2012), https://lawcat.berkeley.edu/record/1125068/files/fulltext.pdf.

> Argues that private governance of platforms is more efficient than public law because private enforcers can better monitor and expeditiously deal with bad behavior, suggesting a need for caution in the application of competition laws to such platforms.

Nikolas Guggenberger, *Essential Platforms*, 24 Stan. Tech. L. Rev. 237 (2020) [hereinafter Guggenberger, *Essential*], https://law.stanford.edu/wp-content/uploads/2021/05/publish_this_-_guggenberger_essential_platforms_eic.pdf.

> Calls for reviving and expanding the essential facilities doctrine to reach digital platforms, thereby requiring them to provide their competitors with "fair and equal access to essential digital platforms."

Andrei Hagiu et al., *Should Platforms Be Allowed to Sell on Their Own Marketplaces?*, 53 RAND J. Econ. 297 (2022), https://onlinelibrary.wiley.com/doi/full/10.1111/1756-2171.12408.

> Based on economic modeling, finds that banning marketplace platforms from selling products decreases consumer welfare, and that a preferable alternative would be to prohibit the platform from close imitation of third-party products or from self-preferencing.

Lina M. Khan, *The Separation of Platforms and Commerce*, 119 Colum. L. Rev. 973 (2019), https://columbialawreview.org/wp-content/uploads/2019/05/Khan-THE_SEPARATION_OF_PLATFORMS_AND_COMMERCE-1.pdf.

Argues that the potential hazards of integration by dominant tech platforms warrant structural separations that preclude dominant platforms from entering lines of business operated on their platforms.

John M. Newman, *Antitrust in Attention Markets: Definition, Power, Harm* (Univ. of Mia., Research Paper No. 3745839, Jan. 10, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3745839.

In view of discrepancies between traditional antitrust rules and zero-price attention markets, proposes alternative approaches for enforcing antitrust law against attention-based platforms.

Interoperability

Herbert Hovenkamp, *Antitrust Interoperability Remedies*, 123 Colum. L. Rev. F. 1 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4035879.

Considers the role of interoperability remedies as a remedy for dominant firms that violate the antitrust laws, in view of different business models, markets, and technological environments.

Thomas E. Kadri, *Digital Gatekeepers*, 99 Tex. L. Rev. 951 (2021), https://texaslawreview.org/wp-content/uploads/2021/04/Kadri.Printer.pdf.

Questions cyber-trespass laws that give platforms property-like rights to exclude, calling instead for a regulatory framework for interoperability and platform data collection.

Mike Masnick, Protocols, Not Platforms: A Technological Approach to Free Speech (Knight First Amend. Inst. at Columbia Univ. 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

Contends that interoperability among platforms, through development of open protocols for social media activities, are a worthwhile approach to contemporary content moderation difficulties such as disinformation, hate speech, and trolling.

Peter Swire, *The Portability and Other Required Transfers Impact Assessment: Assessing Competition, Privacy, Cybersecurity, and Other Considerations*, 6 Geo. L. Tech. J. 57 (2022), https://papers.ssrn.com/abstract=3689171.

> Considers tradeoffs of data portability and related regimes, to identify a framework for assessing such regimes' impact on competition, innovation, user autonomy, consumer interests, privacy, and security.

Moderation Obligations

Niva Elkin-Koren et al., *Social Media as Contractual Networks: A Bottom Up Check on Content Moderation*, 107 Iowa L. Rev. 987 (2022), https://ilr.law.uiowa.edu/print/volume-107-issue-3/social-media-as-contractual-networks-a-bottom-up-check-on-content-moderation/.

> Argues that "contractual networks," an emergent source of legal obligations arising from an interrelated network of bilateral arrangements, provide a basis for challenging platforms' unilateral content moderation decisions.

Quinta Jurecic, *The Politics of Section 230 Reform: Learning from FOSTA's Mistakes*, Brookings (Mar. 1, 2022), https://www.brookings.edu/research/the-politics-of-section-230-reform-learning-from-fostas-mistakes/.

> In view of limited utility and unintended consequence of the 2018 FOSTA amendment to section 230, urges caution for contemporary proposals to amend the same law.

John S. & James L. Knight Found. & Gallup, Inc., Media and Democracy: Unpacking America's Complex Views on the Digital Public Square (Mar. 9, 2022), https://knightfoundation.org/reports/media-and-democracy/.

> Survey of Americans finding mixed views but increasing support for regulation of content moderation.

Olivier Sylvain, *Platform Realism, Informational Inequality, and Section 230 Reform*, 131 Yale L.J.F. 475 (2021), https://www.yalelawjournal.

org/forum/platform-realism-informational-inequality-and-section-230-reform.

Argues that platforms' economic incentives and access to user data demand a new approach to section 230 consistent with consumer protection and civil rights.

Marcelo Thompson, *Beyond Gatekeeping: The Normative Responsibility of Internet Intermediaries*, 18 Vand. J. Ent. & Tech. L. 783 (2016), https://scholarship.law.vanderbilt.edu/jetlaw/vol18/iss4/4.

Argues for intermediary liability to be based on insufficiency of reasoning about content moderation rather than insufficiency of outcomes.

Eugene Volokh, *Treating Social Media Platforms like Common Carriers?*, 1 J. Free Speech L. 377 (2021), https://papers.ssrn.com/abstract=3913792.

Proposes treating delivery of user-generated content, but not curation or algorithmic selection of such content, as common carriage limiting platforms' ability to discriminate on what to take down.

Section 230

Enrique Armijo, *Reasonableness as Censorship: Section 230 Reform, Content Moderation, and the First Amendment*, 73 Fla. L. Rev. 1199 (2021), http://www.floridalawreview.com/2022/reasonableness-as-censorship-section-230-reform-content-moderation-and-the-first-amendment/.

Critiques proposals that require "reasonable" content moderation policies in view of common law interpretations of reasonableness, to conclude that such proposals would be dubious and speech-averse.

Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 Pepp. L. Rev. 9 (2009), https://digitalcommons.pepperdine.edu/plr/vol36/iss2/9.

In view of the increasing importance of technological infrastructure to free expression, argues that the key determinants of free speech's future will be technological design regulation, information commodification, and telecommunications law.

Ashutosh Avinash Bhagwat, *Do Platforms Have Editorial Rights?*, 1 J. Free Speech L. 97 (2021), https://papers.ssrn.com/abstract=3882984.

> Considers the degree to which online platforms enjoy First Amendment editorial rights enabling them to control what content is available on them.

Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. Rev. 713 (2023), https://www.bu.edu/bulawreview/files/2023/10/CITRON.pdf.

> Proposes limiting section 230 immunity for moderation decisions, to platforms that have taken "reasonable steps" to address harmful content.

Ellen P. Goodman & Ryan Whittington, Section 230 of the Communications Decency Act and the Future of Online Speech (German Marshall Fund of the U.S., Policy Paper No. 20, Aug. 19, 2019), https://www.gmfus.org/news/section-230-communications-decency-act-and-future-online-speech.

> Reviews contemporary proposals for amending section 230.

Tim Hwang, *Dealing with Disinformation: Evaluating the Case for Amendment of Section 230 of the Communications Decency Act*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442, *in* Social Media and Democracy 252 (Nathaniel Persily & Joshua A. Tucker eds., 2020).

> Proposes dealing with disinformation campaigns by amending section 230 based on principles of "unfair competition in the marketplace of ideas."

International Human Rights

Aliaa Almehdar, *Freedom of Expression on Social Media Platforms: Facebook's Moderation Behavior on Palestine's May 2021 Movement*, 54 N.Y.U. J. Int'l L. & Pol. 207 (2021), https://www.nyujilp.org/wp-content/uploads/2022/02/NYUJILP_Vol54.1_Almehdar_207-219.pdf.

> Characterizes Facebook's 2021 actions against Palestinian influencers as a clash between free expression and human rights on the one hand, and market and government pressures on the company on the other.

Yohannes Eneyew Ayalew, *From Digital Authoritarianism to Platforms' Leviathan Power: Freedom of Expression in the Digital Age Under Siege in Africa*, 15 Mizan L. Rev. 455 (2021), https://www.ajol.info/index.php/mlr/article/view/220216.

Describes how both African governments' "digital authoritarianism" and platforms' content moderation decisions in Africa.

Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, 38 Yale J. on Regul. Bulletin 86 (2020), https://openyls.law.yale.edu/bitstream/handle/20.500.13051/5440/Benesch._Bulletin._Macro._Final.pdf?sequence=2.

Explores applicability and appropriateness of international human rights law as an influence on social media companies' decisionmaking.

Evelyn Douek, *The Limits of International Law in Content Moderation*, 6 UC Irvine J. Int'l Transnat'l & Compar. L. 37 (2021), https://scholarship.law.uci.edu/ucijil/vol6/iss1/4/.

Critiques the use of international human rights law as a basis for content moderation, arguing that it is ineffective as a constraint on unilateral platform action.

Nicolas Suzor et al., *Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online*, 11 Pol'y & Internet 84 (2019), https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.185.

Proposes use of U.N. Guiding Principles on Business and Human Rights to develop recommendations for platforms to address online gender-based violence.

Transparency

Ángel Díaz & Laura Hecht-Felella, Double Standards in Social Media Content Moderation (Brennan Ctr. for Just. May 20, 2022), https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation.

Argues that platform content moderation policies conceal substantial discretion, and so calls for mandatory transparency particularly around hate speech and harassment removal and independent audits.

Daphne Keller, *User Privacy vs. Platform Transparency: The Conflicts Are Real and We Need to Talk About Them*, Stan. Ctr. for Internet & Soc'y (Apr. 6, 2022), https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0.

> Observes unanswered questions about privacy resulting from legal measures that require platforms to engage in transparency reporting or to give researchers access to internally-held information.

Caitlin Vogus & Emma Llansó, Making Transparency Meaningful: A Framework for Policymakers (Dec. 14, 2021), https://cdt.org / insights / report - making - transparency - meaningful - a - framework-for-policymakers/.

> Taxonomizes transparency measures relating to content moderation into categories of (i) aggregated data on moderation actions, (ii) user notifications about particular decisions, (iii) researcher access to data, and (iv) third party audits.

Platform Governance

Jenny Fan & Amy X. Zhang, *Digital Juries: A Civics-Oriented Approach to Platform Governance*, *in* 2020 Proc. CHI Conf. on Hum. Factors Comput. Sys., https://dl.acm.org/doi/abs/10.1145/3313831.3376293.

> Proposes and evaluates digital juries as an alternative approach to resolution of content moderation disputes.

Access Now, Santa Clara Principles on Transparency and Accountability in Content Moderation: Open Consultation Report (Elec. Frontier Found. 2021), https://santaclaraprinciples.org/images/SantaClara_Report.pdf.

> Identifies baseline principles for platform content moderation regarding publication of aggregate statistics, notice to affected users, and appeal rights.

Rory Van Loo, *Federal Rules of Platform Procedure*, 88 U. Chi. L. Rev. 829 (2021), https://lawreview.uchicago.edu/publication/federal-rules-platform-procedure.

Identifies a need for external rules of content moderation procedure to limit private discretion and enable separation of platforms' executive, legislative, and judicial powers.

Privacy and User Data

Solon Barocas & Karen Levy, *Privacy Dependencies*, 95 Wash. L. Rev. 555 (2020), https://digitalcommons.law.uw.edu/wlr/vol95/iss2/4.

Taxonomizes "privacy dependencies," namely situations in which an individual's private information may be revealed based on third party activities over which the individual may lack control, which challenge individual-based notice-and-consent models of privacy regulation.

Salome Viljoen, *A Relational Theory of Data Governance*, 131 Yale L.J. 573 (2021), https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance.

Argues that personal data is relational rather than individual, so that harms from datafication materialize unjust social relations.

Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 Nw. U. L. Rev. 357 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921003.

Argues that privacy governance is a network of organizational relationships to manage, and not merely a set of data flows to constrain.

Reuben Binns & Elettra Bietti, *Dissolving Privacy, One Merger at a Time: Competition, Data and Third Party Tracking*, 36 Comput. L. & Sec. Rev. 105369 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3269473.

Argues that third-party tracking firms pose unique problems for competition law because mergers between them can create concentrations of data that are not captured by standard antitrust descriptions of the relationship between firms and consumers.

Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. Rev. 793 (2022), https://www.bu.edu/bulawreview/files/2022/04/CITRON-SOLOVE.pdf.

> Argues for relaxation of the harm requirements that typically attend privacy-based causes of action, and classifies privacy harms that could be applied in such cases.

Daniel J. Solove, *The Limitations of Privacy Rights*, 98 Notre Dame L. Rev. 975 (2023), https://papers.ssrn.com/abstract=4024790.

> Argues that effective privacy protection must go beyond individual privacy rights and toward a larger architecture for managing data collection, processing, and transfer, given the complexity and interrelatedness of individuals' privacy interests.

Rory Van Loo, *Privacy Pretexts*, 108 Cornell L. Rev. 1 (2022), https://www.cornelllawreview.org/wp-content/uploads/2022/12/Van-Loo-article-PDF-for-online.pdf.

> Identifies platforms' use of privacy as a pretext for limiting competition, digital helpers, and regulators, which ultimately undermine data privacy's ethos of protecting individuals.

Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019), https://www.publicaffairsbooks.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/.

> Explores the history, mechanisms, and political and individual implications of surveillance capitalism, defined as the process through which major tech-corporations collect and analyze data from users' personal lives and experiences to better predict and market products.

Encryption and Law Enforcement Access

Dhanaraj Thakur et al., Outside Looking in: Approaches to Content Moderation in End-to-End Encrypted Systems (Aug. 12, 2021), https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/.

> Evaluates approaches to content moderation in end-to-end encrypted systems, with focus on user reporting, traceability of message senders,

metadata analysis, and automated client-side detection of problematic content.

Outside the United States

Evelyn Douek, *Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning out Speech*, 94 Australian L.J. 41 (2020), https://papers.ssrn.com/abstract=3443220.

Argues that Australia's mandatory content moderation law overestimates the technological capacity of platforms, incentivizes over-removal, and fails to address underlying social problems.

William Echikson & Olivia Knodt, Germany's NetzDG: A Key Test for Combatting Online Hate (Ctr. for European Pol'y Stud., Research Report No. 2018/09, Nov. 22, 2018), https://www.ceps.eu/ceps-publications/germanys-netzdg-key-test-combatting-online-hate/.

Reviews effect of German content moderation law, finding that it has not led to draconian consequences but also has not caused significant changes in platform behavior.

Francis Fukuyama & Andrew Grotto, *Comparative Media Regulation in the United States and Europe*, https://www.cambridge.org/core/books/social-media-and-democracy/comparative-media-regulation-in-the-united-states-and-europe/0E4F255ADA3FC81BDC4365FF10DFDF3A, *in* Social Media and Democracy, *supra*, at 199.

Compares recent social media regulation developments in the United States, France, and Germany, arguing that most new regulations are effective extensions of already-existing practices.

Emily Irwin & Niloufer Selvadurai, *Imposing Liability on Online Intermediaries for Violent User-Generated Content: An Australian Perspective*, 28 Rich. J.L. & Tech. 1 (2021), https://jolt.richmond.edu/files/2021/11/Irwin-Selvadurai-Final-Verison-.pdf.

Critiques Australian law requiring removal of "abhorrent violent material," questioning whether it may lead to over-removal of lawful or beneficial speech on social media.

João Quintais & Sebastian Felix Schwemer, *The Interplay Between the Digital Services Act and Sector Regulation: How Special Is Copyright?*, 13 European J. Risk Regul. 191 (2022).

Compares interactions between the European Digital Services Act and Europe's regime for online copyright enforcement.

Wolfgang Schulz, *Regulating Intermediaries to Protect Privacy Online—The Case of the German NetzDG*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3216572, *in* Personality and Data Protection Rights on the Internet (Marion Albers & Ingo Wolfgang Sarlet eds., 2022).

Evaluates Germany's content moderation law from perspectives of content over-removal and human rights interests.

Thomas Wischmeyer, *What Is Illegal Offline Is Also Illegal Online: The German Network Enforcement Act 2017*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3256498, *in* Fundamental Rights Protection Online (Bilyana Petkova & Tuomas Ojanen eds., 2020).

Considers whether German content moderation law adequately allocates responsibility for preventing harmful speech among intermediaries, the administration, and the courts.

Patrick Zurth, *The German NetzDG as Role Model or Cautionary Tale? Implications for the Debate on Social Media Liability*, 31 Fordham Intell. Prop. Media & Ent. L.J. 1084 (2020), https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1782&context=iplj.

Proposes Germany's 2017 law requiring blocking of harmful content as a model for U.S. reform of intermediary immunity under section 230.