4

The Problem of Platform Self-control

The platform companies knew, or should have known, that their platforms were being abused during the Trump administration. Unlike in Myanmar, the constant sharing of misinformation and incitement of hate associated with Donald Trump was going on in English, and in the United States – the language and the country best covered by the content moderation tools of every company. It is well established that Facebook in particular realized at the latest shortly after the 2016 election that a small army of trolls, partially in the pay of Russia, was working in favor of Donald Trump (Madrigal 2017).

In response to this knowledge, according to whistleblower Frances Haugen, Facebook undertook measures during 2020 to protect the public from the spread of particularly dangerous information, such as false claims that the election had been stolen (Looft and Ferris 2021). Unfortunately, as Haugen reports, those measures were removed prior to January 6 and were not reinstated until after the coup attempt (Looft and Ferris 2021).

To some extent, the failure to control the spread of election-related lies simply resulted from challenges with enforcing company policies at scale. For example, according to a leaked internal Meta document, events moved too swiftly for company investigative personnel to identify that the various "stop the steal" and related groups were working in concert (Mac, Silverman, and Lytvynenko 2021). Since Facebook is the largest social media company and accordingly had the most resources to do this work, it's hard to imagine that any other company could do better.

And yet, technical and investigatory challenge is not the full story, for there's also plenty of evidence that the company was scared away from addressing right-wing misinformation. During the Trump administration, for example, it was clear that far-right "news" publisher Breitbart was a major source of misinformation. According to another whistleblower, when company employees asked why nothing was being done, Joel Kaplan, the head of Facebook's policy team, replied: "Do you want to start a fight with Steve Bannon?" (Timberg 2021).

I claim that this fear of retaliation by powerful political figures, combined with the temptation to short-term profit and some degree of internal political conflict,

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

drives a problem of self-control that can be seen across the platform context. This chapter aims to offer some solutions for it.

4.1 MARK ZUCKERBERG ISN'T PLOTTING TO FIX THE ELECTION FOR YOUR POLITICAL ENEMIES, I PROMISE

On May 16, 2020, the then-President of the United States alleged – ironically, on Twitter itself – that "The Radical Left is in total command & control of Facebook, Instagram, Twitter and Google. The Administration is working to remedy this illegal situation. Stay tuned, and send names & events."[1] A year beforehand, Trump had alleged that Twitter "make[s] it very hard for people" to follow him, and that "[i]f I announced tomorrow that I'm going to become a nice liberal Democrat, I would pick up five times more followers."[2] According to anonymous sources, Department of Justice officials in the Trump administration were toying with proposals to hold social media companies liable for alleged political censorship before he left office.[3]

Around the same period, Senator Josh Hawley (R-MO) proposed an "Ending Support for Internet Censorship Act," which, in the words of the press release announcing the bill, "removes the immunity big tech companies receive under Section 230 unless they submit to an external audit that proves by clear and convincing evidence that their algorithms and content-removal practices are politically neutral"; the bill was allegedly in response to "a growing list of evidence that shows big tech companies making editorial decisions to censor viewpoints they disagree with."[4] A month after Hawley, Representative Paul Gosar (R-AZ) proposed a similar bill.[5]

These allegations and proposals should be puzzling to a student of American capitalism. After all, social media companies are under a fiduciary obligation to try to produce profits for their shareholders. Organizing themselves on the basis of ideology would alienate – expensively – many of their users, plus the government in any Republican administration in the United States and similar regimes in many other countries. There is a reason that we don't see many US publicly traded companies going all-in on a particular political party. Unsurprisingly, Facebook, Twitter,

[1] Donald J. Trump Twitter May 16, 2020, https://twitter.com/realDonaldTrump/status/1261626674686447621.

[2] Margaret Harding McGill and Cristiano Lima, "White House to Hold Social Media Summit amid Trump Attacks," Politico, June 26, 2019, www.politico.com/story/2019/06/26/white-house-social-media-summit-1383280.

[3] Tony Romm, "Attorney General Barr Blasts Big Tech, Raising Prospect That Firms Could Be Held Liable for Dangerous, Viral Content Online," *Washington Post*, February 19, 2020, www.washingtonpost.com/technology/2020/02/19/attorney-general-barr-blasts-big-tech-questioning-its-protection-liability-content/.

[4] "Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies," June 19, 2019, www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies; S. 1914.

[5] H.R. 4027; "Congressman Gosar Introduces Legislation to Stop Big Tech Censorship," July 25, 2019, https://gosar.house.gov/news/documentsingle.aspx?DocumentID=3854. For a description of numerous other efforts to amend or repeal Section 230, see Samuelson (2021).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

and YouTube deny that there is any political censorship going on their platforms (with the exception of some Musk claims about prior Twitter practice), and no partisan criteria can be discerned anywhere in their published content rules. Thus, conservatives in the United States who accuse them of censorship typically do not accuse them of having explicit and formal partisan policies – mainstream conservatives do not claim that there's a "no Republicans" rule somewhere in Facebook's Community Standards. Rather, the basic theory seems to be something like that company employees – located, after all, primarily in notoriously liberal Northern California – are bringing their own political ideologies to their policy enforcement decisions in the context of ambiguous rules and nontransparent enforcement systems (e.g., McGregor and Kreiss 2019).[6]

But this isn't just some kind of right-wing conspiracy theory. The left makes similar allegations, and they are just as puzzling, if slightly different. The left, rather than accusing the companies of over-enforcing their content-moderation rules in a politically biased way, accuses them of under-enforcement against right-wing violations of those rules. For example, then-Senator Kamala Harris sent a letter to the Twitter CEO recounting several of Trump's tweets, alleging that they violate the Twitter terms of service, and seeking the removal of Trump's account more than a year before his account got removed for the January 6 autogolpe bid.[7] Persistently, commentators on the left have accused the companies of running scared from threats of regulation by the right – or simply of being more interested in the profits to be gained from the engagement that people like Donald Trump generate than in the safety of their other users and the country – and hence failing to control harassment, hate speech, and dishonest propaganda that violates their own policies.[8]

The left-wing critique makes marginally more economic sense, at least in the short-term, than the right-wing one. Conceivably, a racist or violent political leader could drive enough profitable engagement that it might be in the interest

[6] Perhaps there is also a secondary claim that some platform rules themselves, such as prohibitions on hate speech, are inherently biased. This claim isn't made so frequently or loudly, I guess because it's still considered impolite to admit that one's political ideology entails racial slurs and the like.

[7] Letter from Kamala Harris to Jack Dorsey, October 1, 2019, https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/r3ukxGLqQFLA/v0 (last visited March 14, 2020) (see also Ziady 2019). Twitter has carved out a formal exception for people like Trump. Twitter, "About Public-Interest Exceptions on Twitter," https://help.twitter.com/en/rules-and-policies/public-interest [https://perma.cc/GA5Q-FG2C] (last visited December 4, 2022), describes an exception to the Twitter rules for policy violating content by elected officials, and says that such content will be put "behind" a "notice" of its rule violating character.

[8] Even after Trump was banned, some commentators suggested that Facebook's particular approach to the ban was motivated by advertising interests. For example, Ryan Goodman of NYU, commenting on the Oversight Board's upholding of Trump's Ban: "Did Facebook have financial incentive to avoid permanently banning Trump? This in @OversightBoard decision struck me: Facebook refused to answer Board's question 'whether account suspension or deletion impacts the ability of advertisers to target the accounts of followers'." post of May 5, 2021; https://twitter.com/rgoodlaw/status/1389954225409601538.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

of Facebook's or Twitter's shareholders to leave their hate speech alone. But that is true only if the leaders of those companies heavily discount the future. After all, it's no coincidence that Google's content regulation team is called "trust and safety": A social media platform filled with hate speech, threats of violence, or harassment is likely in the long term to drive away its users in fear or horror, either to competitors or to no platform at all.[9] (For reasons noted later in this chapter, this user flight is likely to come by surprise and all of a sudden.) Platform leaders know that, so the left-wing critique seems to assume that they're willing to sacrifice the long-term security of their user-base for a few extra clicks in the immediate present. Yet these same leaders have a well-known predilection to make very large, very long-term investments, such as Alphabet's investment in its famous "moonshot" division, X, and Facebook's creation of "Internet.org" (thereafter "Free Basics") to expand internet connectivity in developing markets, plus its huge acquisitions made at a staggering cost in order to increase the number of users it can reach without an immediate revenue source, most famously its acquisition of WhatsApp. This is not the behavior of corporate executives with a short-sighted focus on the next quarter's profit and loss statements.

The potential flight of their users is not the only way that toxic content harms platform interests. As Gillespie (2018a, 14) recounts, there has at least been (plausible) speculation that the toxicity of Twitter's environment scared off possible corporate acquirers years before Elon Musk bought it. Moreover, many advertisers, concerned with "brand safety" – that is, with avoiding the risk of associating their brands with upsetting or disreputable content – are likely to flee platforms with too high a proportion of toxic content, even if adherents to one of the country's major political persuasions find that content less objectionable (Bellman et al. 2018; Kulp 2019; Braun and Eklund 2019, 6; for a chronology of YouTube's brand safety challenges, see Pottinger 2018–2019, 525–31).

In short, neither the censorship of innocent speech nor the refusal to remove harassment and propaganda is likely to be consistent with the long-run economic interests of social media companies. If those companies can actually control their own behavior, both the left-wing and the right-wing complaints fail the test of plausibility. Yet, as I will discuss further in a moment, there is a little bit of evidence that some of the complained-of behavior is actually happening. How? I claim that, to the extent either complaint has any grounding in fact, what really has to be going on is a kind of failure of *corporate self-control*, a succumbing to short-term temptation to squeeze out a few extra clicks or satisfy a few noisy left-wing engineers or vengeful politicians at the expense of long-run company interests.

There is some evidence supporting the complaints of both sides in the US political context. For example, one detail prominent in many of the right-wing complaints

[9] For an interesting recent economic analysis of the incentives underneath social media content moderation, see Liu, Yildirim, and Zhang (2021).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

is a practice known as "shadowbanning," in which a user isn't actually banned from the platform, but the content they produce is silently made invisible, or less visible, to other users. And while the companies deny political shadowbanning, it is uncontested that they have, and exercise, the power to shadowban – they admit to using reductions in the distribution of content as a lever for enforcing their policies.[10] Moreover, both employees of technology companies and wealthy technology entrepreneurs really do predominantly lean to the left, or, at least, support Democrats over Republicans (Broockman, Ferenstein, and Malhotra 2019). There was also at least one relatively concrete and high-profile case that could fairly be ascribed in part to political bias in content moderation on Twitter, where the company refused to allow the distribution of a New York Post story about Hunter Biden in what appears (though much is still unclear) to be an erroneous application of company rules about content derived from hacking (Vlamis 2022; Schreckinger 2022).[11]

Similarly, the left-wing complaint also benefits from some factual support: There are some credible press reports suggesting that Facebook officials in fact consciously interpreted their platform rules charitably toward conservatives in order to avoid an appearance of left-wing bias or offending conservative lawmakers.[12] More generally, there is evidence that at least some platform company executives were in fact motivated by short-term growth metrics as opposed to longer-term platform integrity. A whistleblower report to the SEC, FTC, and DOJ by Peiter Zatko ("Mudge") alleges that Twitter executives turned off measures meant to prevent "spam bots" because they were rewarded for increasing measures of active users.[13] The economics of platforms can undermine rule enforcement in other ways. For example, in some

[10] See, e.g., Facebook Community Standards #20: False News, www.facebook.com/communitystandards/false_news (last visited March 15, 2020), which states that "we don't remove false news from Facebook, but instead significantly reduce its distribution by showing it lower in the News Feed"; Twitter, "Our range of enforcement options," https://help.twitter.com/en/rules-and-policies/enforcement-options (last visited March 15, 2020), which include "[l]imiting Tweet visibility." Gillespie (2022) further describes shadowbanning and related phenomena. It's worth nothing that shadowbanning is the obviously correct solution in some cases. For example, the dating site Bumble reportedly uses a form of shadowbanning to attempt to get rid of users accused of sexual assault – from the user's perspective, their profile is still on the site, but nobody else can see it, hopefully making it harder for the "silent block[ed]" user to know they're banned and create another account or retaliate against the person who reported them (Edwards et al. 2021). Given the extreme harms inflicted by sexual assault, this seems like an entirely reasonable policy to me (even done on the basis of mere accusation, since an alleged assaulter's interest in being on a dating app isn't significant enough to be entitled to any substantial process before acting proactively to protect those with whom they might interact).

[11] Well-known technology industry blogger Mike Masnick has an overall summary of the events surrounding the Hunter Biden laptop leak at "Hello! You've Been Referred Here Because You're Wrong about Twitter and Hunter Biden's Laptop" (TechDirt, December 7, 2022), www.techdirt.com/2022/12/07/hello-youve-been-referred-here-because-youre-wrong-about-twitter-and-hunter-bidens-laptop/.

[12] Journalist Steven Levy (2020a, 340–45), who was given an unusual amount of access to Facebook personnel, reports on a number of instances in 2015–6 where Facebook was particularly solicitous to conservative fears of "censorship" even of rule-violating content.

[13] Whistleblower report of Peiter Zatko, https://s3.documentcloud.org/documents/22186782/whistleblower_disclosure.pdf, pp. 10–12.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

companies the "adjudication" of rule violation is merged with a customer service function, creating inconsistent rule enforcement across revenue classes. Persistent public controversy about Meta's "crosscheck" system, which allegedly provided for special scrutiny before platform sanctions could be levied against prominent (and hence revenue generating) accounts is one example (Horwitz 2021). According to the Oversight Board's evaluation of that system, not only figures of public importance such as elected officials are included – so are "business partners" who are relevant to Meta's bottom line.[14] Another example is YouTube's tiered system of customer service for content creators, which Caplan and Gillespie (2020, 7) allege provides greater affordances to appeal platform sanctions for higher revenue users.

Again, I think the evidence is somewhat stronger for the left-wing complaint (though this evaluation may be colored by my own left-wing politics). It is consistent with evidence of similar company behavior outside of the country: There are credible allegations that Meta also soft-pedaled rule enforcement against the Modi regime in India (Purnell and Horwitz 2020). Allegations about Meta's failures in India are strikingly parallel to those in the United States: In both countries, a high-level company "policy" official – where "policy" is an organizational function simultaneously responsible for lobbying and government relations and for rule development and enforcement – allegedly intervened in rule enforcement in order to protect a government in power, with allegedly mixed motives to both promote that official's personal ideology aligned with the government and to shield the company from government retaliation.[15] One person came up with an experiment to tease out the inconsistency in Facebook policies: A Facebook user created a page entitled "Will they suspend me?," which quoted Trump's posts to see whether the same standard would be applied to an ordinary person (O'Kane 2020). The same standard was not applied.

Even fairly small platforms face serious problems with consistent rule enforcement. In 2021 OnlyFans, the amateur subscription video platform that became a

[14] Oversight Board policy advisory opinion on cross-check program, December 2022, https://oversightboard.com/attachment/440576264909311/, p. 8. See also ibid., p. 30, for an example, in which a celebrity appears to have been given favorable content moderation treatment through the cross-check system in anticipation of signing an exclusive deal with a Facebook streaming service.

[15] Compare the press accounts of the behavior of Ankhi Das (Purnell and Horwitz 2020) and Joel Kaplan (Mac and Silverman 2021). Another press report suggests that there may have been similar dynamics in Brazil and in the U.K. In Brazil, a "subject-matter expert" refused to permit the removal of a speech by Jair Bolsonaro in which he described indigenous Brazilians as "evolving and becoming, more and more, a human being like us" (Marantz 2020). But it turned out that the "expert" in question had previously worked for a political ally of Bolsonaro's. The employee who spoke to the press suggested that the refusal to remove the speech was likely also motivated by advertising revenue. That same report described far-right pages in the U.K. that were "shielded" – that is, excluded from ordinary rules requiring a ban of a page after enough content policy violations. According to the article, content moderators perceived that shielded pages "tended to be those with sizable follower counts, or with significant cultural or political clout – pages whose removal might interrupt a meaningful flow of revenue" (Marantz 2020).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

center of the online sex worker trade during the COVID-19 pandemic, announced that it would ban pornography (as far as I can tell, the only thing that anyone has used it for). Unsurprisingly, this ban didn't last – less than a week later someone realized that doing so would destroy the company and found some way to make keeping the pornography work (Spangler 2021). But while the stated reason for the short-lived porn ban on the porn site involved pressure from payment processors and other financial intermediaries, one might reasonably suspect that some of that pressure to stop hosting pornography was related to evidence turned up in a BBC investigation that it had failed to enforce its own rules. It turns out that OnlyFans had been hosting unlawful conduct (in relevant jurisdictions) such as prostitution and, most alarmingly, sexual videos featuring people under 18 (Titheradge 2021). Making matters worse, the BBC reported (Titheradge 2021) that more popular (and hence profitable) accounts were given – as a matter of written policy! – more "warnings" before being shut down for violations of platform rules, including rules about illegal content. It sure seems like a failure to comply with their own rules or underlying law almost killed the company.

Nonetheless, while supported by some evidence, both the left-wing and the right-wing complaints about the major social media companies are almost certainly exaggerated. Some complaints may simply be due to differences in interpretation of platform rules; so long as companies are enforcing their rules in good faith, such differences in opinion ought not to be seen as unfair political bias. For example, Kamala Harris's letter to Twitter accuses Trump of violating Twitter rules prohibiting "harassment" and "the glorification of violence," and cites as examples tweets in which Trump intimated that a whistleblower in the executive branch was guilty of espionage. Arguably, those tweets, and the others that Harris cited (including a particularly menacing one about "a Civil War like fracture in this Nation" if Trump was successfully removed from office by impeachment) constituted harassment or threats of violence. In Harris's words, "These tweets should also be placed in the proper context, where the President has compared the whistleblower to a 'spy' who may have committed treason, and further implied that the punishment for that should be death."

However, an equally reasonable person could believe that the tweets were no such thing. "Harassment" is a notoriously slippery concept. And Trump's tweets, regardless of what one might think about their overall democratic propriety, could just as easily be interpreted as nothing more than the attempts of a politician facing serious accusations to defend himself by impugning the behavior of his accusers and by warning of the political (not violent) consequences of their actions. Similarly, the "Will they suspend me" disparity between enforcement against Trump and against an ordinary person can potentially be attributed to the existence of the infamous cross-check system – according to which prominent users weren't subjected to different rules but were afforded different *process* in the form of a second level of company scrutiny before their Facebook posts were taken down. According to the person behind "Will they suspend me," Facebook ultimately claimed that one of his posts was removed in

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

error (O'Kane 2020), and this is consistent with the notion that Facebook was less willing to accept erroneous removals of the content of prominent people like Trump.[16]

The examples of Harris's letter and "Will they suspend me" suggest what is intuitively obvious to most of us as I write these words at the end of 2022: The social media platforms are faced with pervasive distrust and skepticism relating to the fact that their rules are not enforced transparently and are also subject to substantial debate in their application. But faithful enforcement in accordance with platforms' own rules may be just as damaging to their bottom lines as inconsistency if the external world cannot observe that enforcement actually is consistent. If platforms cannot reassure users and regulators that they are neutrally enforcing their rules, then they might suffer a lack of user trust or retaliatory regulation regardless of the actual state of affairs inside corporate offices. To illustrate this phenomenon, we can note that, regardless of whether pre-Musk Twitter was actually censoring conservatives (e.g., whether that Hunter Biden incident was a good-faith mistake or colored by the partisan affiliation of the company employees who made the call), enough people on the right believed that censorship was happening that many fled to various extreme right wing "free speech" social media platforms such as Parler, Gab, and "Truth Social" (the last of which was founded by Donald Trump). "Gab," for example, competes with other platforms on the basis of its lack of "political censorship" (Zannettou et al. 2018; Lima et al. 2018). Twitter's loss of a sizeable chunk of the users from one political affiliation is a problem from the standpoint of a company whose revenues are tied to scale; it may also be a problem from the standpoint of society to the extent the flight of the far right into unmoderated echo chambers promotes their further polarization and radicalization. Thus, platforms need not only enforce their rules neutrally but also must convince their users to trust that they are doing so. Even if they can control themselves, in other words, they need to be able to control themselves *in public*.

4.1.1 *Sometimes Failures of Self-control Are Just Failures of Corporate Governance*

In some cases, an organization's failures of self-control do not reach its top ranks. Low-level employees might frustrate the policy choices of top-level leaders by disobeying the rules, or by distorting their application. For example, the frontline

[16] Such a policy isn't necessarily irrational or bad – it's problematic if purely motivated by revenue, but it might make sense for public discourse reasons – because of the greater attention paid to posts by particularly prominent people, it might make more sense to be more careful about disruptively removing and then restoring those posts. On the other hand, this argument only holds if the cross-check process happens quickly, otherwise prominent people could – as Trump in effect did – take advantage of the extra time it affords to use their gigantic audiences to do immense public harm. This delay is one of the risks identified by the Meta Oversight Board in its policy recommendations surrounding cross-check, see Oversight Board, "Oversight Board publishes policy advisory opinion on Meta's cross-check program," www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/ (December 2022).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

workers charged with interpreting social media content rules might be interpreting – intentionally or inadvertently – ambiguous rules concerning "harassment" in such a way that conservative speech is seen as harassing at a different rate than liberal speech, independent of the inherent nature of such speech (if such a thing is conceptually coherent).

This is unlikely to be the case with respect to the most politically salient controversies, such as the debates over social media bans of particular prominent conservatives. Low-level employees are not quietly making the decisions about whether or not to ban Trump from Twitter or Alex Jones from Facebook.[17] However, to the extent general patterns of enforcement or nonenforcement are nonetheless skewed in less prominent cases, this could, in the aggregate, undermine both the capacity of these platforms to preserve the environment their leaders are trying to create and public trust in their neutrality.

Another example of the failure of corporate self-control through employee defection comes, not from social media political censorship and hate speech, but from Amazon's marketplace. Amazon owns a number of private label brands, such as "Amazon Essentials," which compete with the products of third-party sellers on its platform. Amazon also collects immense amounts of competitively valuable sales data on behalf of its third-party sellers. This creates an obvious conflict of interest: Amazon has a short-term incentive to effectively engage in industrial espionage against its own sellers. In the absence of some mechanism for corporate self-control, this conflict of interest could deter sellers from using Amazon to distribute their products; accordingly, the company has a policy of not using individual seller data from its third-party sellers in choosing which private label products to release. Recently, it came out that Amazon's employees violated this policy, exploiting loopholes in the firewall between the sales data side of the business and the private label side to do so.[18]

Amazon's failure to control employees resembles more general failures of employee management that have allowed companies to stumble into unethical conduct. Familiar cases of such employee malfeasance in the nonplatform economy include foreign sales employees succumbing to the temptation to bribe government officials and mortgage brokers writing "stated income" (a.k.a. "liar") loans. Hence, to some degree, such problems might be amenable to conventional management techniques such as tightening the enforcement of internal rules and conducting random audits.

However, because such internal controls are always imperfect, and Amazon is a platform, it also raises many of the same user trust issues as political "censorship"

[17] Similarly, emails that Musk released from Twitter relating to the Hunter Biden story noted above show the involvement of various senior Twitter personnel at the time, including its head of trust and safety and its general counsel.

[18] The details of this story come from Dana Mattioli (2020). Sam Bowman (2020) of the Adam Smith Institute has a helpful discussion of the incentive Amazon has to avoid misusing third-party seller data in order to avoid deterring product innovation on its platform.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

on social media. Amazon does not only need to control its employees, it needs to make control of its employees credible to third-party sellers. This contrasts with other contexts of employee misconduct – for example, mortgage brokerages do not need to credibly signal to their customers that they have adequate internal controls to stop writing liar loans.[19] This illustrates that when employee and leadership goals diverge, traditional techniques of management do not exhaust the strategies of corporate self-control that are relevant; when there are external constituencies who need credible signals of trustworthiness to participate in a platform ecosystem, companies also require tools to make themselves externally accountable.

4.2 CAN PLATFORMS COMMIT THEMSELVES TO GOVERN CONSISTENTLY?

Lack of self-control doesn't come out of nowhere. We have fairly ordinary ways of thinking about its sources. For example, an organization may lack internal management capacity: Corporate leaders might issue directives but lack the power to make the subordinates who actually have their fingers on the metaphorical button that deletes a particular piece of content or bans a particular user comply with them. Or an organization's leadership might suffer from a lack of willpower: They might really want to act in a long-term profit-maximizing way but succumb to the temptation to leave up Alex Jones's conspiracy theory or delete some Republican's speech in the face of short-term profit or political pressures.

The problem of top-level leadership is worth further examination. Right now, the dominant framing of the problem of excessive power in social media content moderation is one of excessive company power. On this framing, companies are understood as monolithic entities with a unified will – understood, for example, from the American political right, to mostly be represented by the general political ideology of Northern California, hence the endless complaints that left-wing technology workers are censoring conservatives.

[19] However, the government may require them to communicate credible information about their internal controls so that it can economize on investigative costs. Also, to some extent mortgage brokers may need to credibly signal to lenders/purchasers of mortgages that they do not write liar loans, but given that there are many fewer mortgage lenders and buyers than there are Amazon sellers, and lenders and buyers are likely to have their own investigative resources and benefit from existing infrastructure such as third-party auditing (which is obviously imperfect, as the financial crisis taught us), the problems are less difficult than they are for a many-to-many platform like Amazon. This dynamic arguably also exists in other non-platform contexts. For example, airlines need to make sure customers trust their employees' compliance with safety regulations – although they have massive government regulation helping them to do so. In some platform contexts, companies might also welcome government regulation in order to have a third-party guarantor of their conduct. Amazon's interest, for example, would probably be served by such a regulation (combined with real auditing and enforcement) insofar as it could then tell sellers "we won't steal your data, because if we do, the government will impose massive fines on us."

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

Yet press reports as well as the accounts of former company employees, for example from the famous "Facebook Files" leak, suggest, to the contrary, that there are routine differences between company employees responsible for rule implementation and senior managers (Birnbaum 2021).[20] And the shape of those differences resembles a familiar problem in political theory: High-level executives are tempted to command deviations from general rules, either to meet short-term crises or to achieve short-term gains against the long-term benefit of rule enforcement.

To illustrate this dynamic, consider the saga of Donald Trump's Twitter and Facebook accounts. During Donald Trump's presidency, there was a widespread public debate about the extent to which his social media posts – like those of other far-right figures – violated platform rules. Part of the problem is that there was a certain degree of ambiguity with respect to those rules in the first place; in particular, at least until April 2018, significant parts of Facebook's content policy enforcement guidelines were not available to the public (Bickert 2018). However, there were certainly plausible arguments that Trump's behavior violated the rules of all the major platforms relating to, *inter alia*, inciting violence, hate speech, and sharing misinformation (arguments that Kamala Harris offered).

There is also a substantial amount of evidence from press accounts of employee complaints at least at Facebook that high-level executives intervened to protect Trump's social media accounts, along with those of other American far-right figures, against rule-enforcement actions that would otherwise have been undertaken by line employees (Mac and Silverman 2021; Dwoskin, Timberg, and Romm 2020; Frier and Wagner 2020; Solon 2020).

Yet high executive power giveth and high executive power taketh away: After the events of January 6, there was strong reason to believe that the political stability of the United States – which happens to contain the headquarters, the vast majority of the regular employees, and probably most of the assets (unless hidden in offshore tax havens) of the major platform companies – was in severe danger. There was a realistic threat of a coup in the United States; the culpability of the social media companies for facilitating its incitement could potentially have been a fatal public relations disaster if it led to mass user or advertiser defection or a severe legislative response.[21] Accordingly, both Facebook and Twitter finally acted, banning Trump from their platforms as perhaps the most prominent part of the series of company actions that has since been going by the name "the great deplatforming."[22] In both cases, credible media reports suggest that the decision was made directly by top-level

[20] There's no particular reason to think this problem is limited to Facebook, it's just that Facebook is the only company that had such a huge leak.

[21] Moreover, if Trump had actually managed to seize authoritarian rule at that moment, how long would Zuckerberg and Dorsey have been allowed to keep their companies, or their freedom?

[22] Because of its infrastructural role, Amazon's removal of the right-wing social networking company Parler from AWS may have been more controversial within the industry.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

executives, that is, Zuckerberg and Dorsey (Byers 2021) – and Dorsey's very public musing on the decision on Twitter certainly is consistent with this.[23]

There's certainly a story that can be told in which the change in Trump's social media status resulted from good-faith interpretations of platform rules in the face of changing circumstances. For example, the interpretations of those rules might be amenable to change in the context of the broader social threat posed by a user's actions.[24] However, it seems much more likely that decisions around Trump were made in an essentially ad hoc fashion at critical points by company leaders both when they kept his most problematic posts up and when they ultimately banned him.

These events thus illustrate both the benefits and the dangers of top-level executive power. On the benefit side: It has long been understood that a key function of executives in political states is to respond to emergencies, and many scholars have suggested that deviating from or suspending the ordinary operation of law in such states is permissible – or at least inevitable – under such circumstances. Carl Schmitt built an entire theory of sovereignty out of this function of executives (which Chapter 5 will discuss at length – I think the Trump situation actually reveals some of the flaws in Schmitt), and it appears in numerous examples of positive law, such as Article 16 of the French Constitution and the United States National Emergencies Act. In view of the extreme danger on January 6, 2021, both to the companies and to the country in which both companies are headquartered, it is reasonable to defend the actions of Zuckerberg and Dorsey – as well as whichever decision makers at Amazon decided to ban Parler, whoever at Reddit decided to ban various Trump-associated subreddits, and so forth – as necessary emergency steps.

On the other hand, however, there's a plausible case to be made that executive power at both companies brought them – and the United States – to that extremity in the first place. Suppose we accept the – controversial but eminently believable – claims that Donald Trump's social media posts routinely violated Twitter and Facebook rules for years beforehand and that Trump's social media activity was necessary (in a causal sense) to the crisis – that is, that the attack on the Capitol would not have occurred in the absence of Trump's capacity to spread lies and incitement over social media to those of his supporters who were most detached from reality. Then we have to conclude that high-level executives caused the very problem that they were forced to solve at the last moment.

This pathological consequence of unconstrained executive power ought not to be surprising. One well-understood feature of agency is that unconstrained

[23] See the Twitter thread starting at https://twitter.com/jack/status/1349510769268850690 (January 13, 2021) and particularly the reference to "the power an individual or a corporation has over a part of the global public conversation" at https://twitter.com/jack/status/1349510772871766020 – one suspects that one knows who the "individual" is.

[24] For example, from Dorsey's January 13 thread: "Offline harm as a result of online speech is demonstrably real, and what drives our policy and enforcement above all." (https://twitter.com/jack/status/1349510770992640001).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

moment-by-moment decision-making capacity actually undermines the autonomy of an agent, whether individual or organizational (Elster 2000). In particular, the inability to bind one's later self to some constraints radically undermines the capacity to make long-term plans or make commitments that are sufficiently credible to permit the making of deals with third parties. Among other things, this fact is a key justification for the existence of the legal form of contract (Fried 2015, 13–14).

Elsewhere, I have analyzed the problem of costly rule-enforcement by states through this lens (Gowder 2016, 59–62). The analysis directly applies to companies as well. It is likely that company executives saw the enforcement of their rules against Donald Trump before January 6, 2021 as costly in the short term, as they were subject to intense political pressure and threats of retaliation by right-wing political leaders over a supposed bias against conservatives (e.g., Leary and McKinnon 2020). And there is evidence from leaked internal memoranda that at least some Facebook employees understood the problem to be that top-level executives, not constrained by their own rules, were vulnerable to short-term political pressure (Hagey and Horwitz 2021).

So – if the foregoing is true – then at least partial blame for the Trump problem in the first place might be laid at the feet of an inability of platform leaders to bind their own decisions. In other words, if Zuckerberg and Dorsey had, circa 2016 or so, the power to bind themselves to enforce their rules without regard to the identity or political or economic power of the rule violator, they could maybe have controlled Trump's behavior long before it posed a threat. And doing so would also have made the companies more robust against retaliatory threats from the right, since the benefit of those threats to their makers would have been less apparent – people like Josh Hawley would have less reason to believe that the companies would back down in the face of more-or-less empty threats of legislative retaliation – potentially moving those threats off the equilibrium path.[25]

The capacity of commitment to increase an agent's resistance to possibly empty external threats is sufficiently important that it's worth filling out in a little more detail. External estimates of the incentives facing Republican lawmakers during the period when they controlled both Houses of Congress and the Presidency carry a substantial amount of uncertainty. They may have sincerely believed that social media companies were biased against conservatives, or they may have merely been saying that in order to stir up anger in their constituents and prime those constituents to disbelieve things like fact-checking of the misstatements of their political allies on the platforms. Moreover, even if Republican lawmakers sincerely believed that the companies were biased, both threatening legislation and actually seriously

[25] "Off the equilibrium path" is a concept from game theory often used in models of threat and deterrence. Speaking informally, we can understand it in the present context as capturing the idea that a player can sometimes rationally commit to an irrationally costly course of action in order to make it irrational for another player to do the thing triggering the committed-to threat. More, including an example, below.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

attempting to enact legislation come with costs – with the costs of a serious attempt somewhat higher – including the expenditure of political capital in deal-making, possible embarrassment if efforts to legislate fail or if enacted legislation ultimately is struck down by the Supreme Court on First Amendment grounds, and even potential national economic loss from damage done to the companies. Even successful legislative efforts have the cost of no longer being able to use social media "censorship" as a political issue. Against these costs must be weighed the advantages of securing additional opportunities to reach social media users with their messages by deterring rule enforcement against their political team.

Depending on the specific weight legislators give to each of those incentives, there are plausible utility profiles in which a legislator would prefer to threaten to regulate social media firms in order to induce them to grant leeway to their team's content without any intention of actually following through with those threats – particularly if they do not actually believe that such firms are enforcing their rules with a bias against their ideological allies (in which case legislation requiring neutrality would likely be ineffective), or if they do not believe that there is a realistic chance of successfully enacting and enforcing such legislation. Judging by the numerous threats that seem to have gone nowhere even in Republican-controlled branches of government during the Trump administration, it seems likely that one of these utility profiles was in play for leading Republicans during that period.[26] But in the face of uncertainty as to whether all these threats are sincere, it was rational for a company executive to put a thumb on the rule-enforcement scale in favor of the group making the threats, that is, American conservatives, to avoid them being carried out.

Under such circumstances, effectively committing a company to enforcing preexisting platform rules against conservative content would at least partially defang

[26] For example, Missouri Republican Josh Hawley introduced S.1914 in June 2019, which proposed to strip Section 230 protection from companies that engaged in "politically biased" content moderation, but the bill appears to have died in a Republican-controlled committee with no action taken. In September 2020, Mississippi Republican Roger Wicker introduced S.4534, entitled the "Online Freedom and Viewpoint Diversity Act," to substantially limit the scope of Section 230 protection, but that bill appears to have died, like Hawley's, in the Commerce, Science, and Transportation committee – of which Senator Wicker was chairman at the time. In the same year, Georgia Republican Kelly Loeffler introduced both S.4062 ("Stopping Big Tech's Censorship Act"), and S.4828 ("Stop Suppressing Speech Act of 2020"), both of which died in the same committee, as did Lindsey Graham's S.5020 to repeal 230 altogether. By then the Democrats controlled the House, but even when the Republicans controlled both branches in the 115th Congress, there was no action on Texas Representative Louie Gohmert's H.R.7363 ("Biased Algorithm Deterrence Act of 2018"). In October 2020, then-FCC chairman Ajit Pai threatened to engage in a rulemaking process on Section 230, but left office having taken no action (Hollister 2021). It's worth noting by way of caveat that the failure of these efforts to gain traction may not be because they weren't sincere threats – they may have been sincere but withdrawn because they successfully induced enough compliance that the Republicans didn't need to follow through. On the other hand, part of the reason that Zuckerberg and Dorsey felt free to act after January 6 may have been because the lack of follow-through served as evidence that the threats were not sincere (or at least were not imminent, the Democrats having taken control of the White House). Florida and Texas Republicans, evidently under different incentives from their federal colleagues, did manage to legislate at the state level.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

the empty threat strategy. There's less reason to make empty threats if their victim cannot surrender to the pressure they create. To be sure, such threats might still be a useful means of voter mobilization (or undermining the credibility of platform fact-checking); however, their incentive and hence their incidence may be reduced to the extent they are motivated at least in part by the capacity to intimidate corporate executives into rule underenforcement.[27]

4.2.1 *Lessons in Self-binding from Political Science*

Political science has a well-developed conceptual apparatus to address these problems of credible commitment (or, depending on the context, sometimes "credible threat"). In the political science context, these ideas appear most prominently in the literature on international relations, relating to the capacity of states to credibly threaten costly military action against one another (e.g., Kilgour and Zagare 1991; Huth 1999), and in the literature on domestic law enforcement, relating to the capacity of states to credibly commit to costly punishment of lawbreakers (e.g., Baker and Miceli 2005). The broad strategic problem is similar across both contexts, so I will simply describe the domestic example, as it is more analogous to the problem faced by platforms.

Consider the following toy problem. A dictator, Caligula, wishes to collect taxes. Naturally, citizens won't pay up voluntarily. Hence, Caligula requires a military/police force to make them do so. However, deploying coercive force is costly – soldiers must eat, ammunition must be acquired, and so forth. Let's suppose that the average cost of punishing a citizen is $1,000. To make life a little easier, we will also assume that the punishment that Caligula can inflict is adequately painful to deter tax evasion, even considering the probability that some tax evaders will not be detected. Unfortunately, Caligula finds that most citizens' tax liability is less than $1,000, and, even if she expropriates all of the assets of every citizen who is found to have evaded their taxes, many citizens' all-in net worth is still less than $1,000. Should someone who is worth less than that amount fail to pay their taxes, it is

[27] A related context may be informative. Network security company Cloudflare terminated the accounts of the Daily Stormer and 8chan because of the vile nature of their content, but ultimately changed their policy to forbid themselves from doing so in large part because exercising such discretion appears to have rendered them vulnerable to external pressure. In the company's words: "In 2017, we terminated the neo-Nazi troll site The Daily Stormer. And in 2019, we terminated the conspiracy theory forum 8chan. In a deeply troubling response, after both terminations we saw a dramatic increase in authoritarian regimes attempting to have us terminate security services for human rights organizations – often citing the language from our own justification back to us." Matthew Prince & Alissa Starzak, "Cloudflare's abuse policies & approach," Cloudflare Blog, August 31, 2022, https://blog.cloudflare.com/cloudflares-abuse-policies-and-approach/. Cloudflare further claimed that "each showing of discretion" in their choices about services to terminate "weakens our argument" in legal challenges to orders seeking to have them carry out global restrictions on, for example, defendants in copyright cases.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

irrational for Caligula to expend the costs necessary to punish them: She'll spend more than she can get back. In the absence of some way to commit in advance to punishing everyone who evades taxes, any person worth less than $1,000 will look down the game tree and realize that they're not in any genuine danger of punishment – so they simply won't pay.

Suppose, however, Caligula can make an irrevocable commitment to punishing tax evaders, no matter how expensive it is? The famous doomsday machine in Dr. Strangelove (a movie much beloved by all game theorists) is the paradigm case – an unstoppable machine set to launch a retaliatory nuclear attack without any human intervention. If Caligula can create an unstoppable tax-enforcement machine – even if that machine still costs $1,000 every time it turns itself on – then even a poor citizen will be aware that the machine will come for him if he fails to pay his taxes. Now, every citizen has an incentive to pay their taxes (remembering our earlier assumption that the punishment is painful enough to deter everyone who genuinely faces its threat). And – the most delightful part for Caligula – because everyone pays their taxes, the punishment machine never turns itself on, and hence she never has to pay the cost of punishing – in the lingo of game theory, tax evasion, and its punishment are "off the equilibrium path."

Thus, credible commitments are the canonical way that a state solves its problem of under-punishing. But a state also needs to refrain from over-punishing. "Over-punishing" in this context means using punishment as a means of expropriation, that is, engaging in revenue-seeking punishment in excess of what is permitted by the law. We often use the language of the rule of law to describe the imperative for states to follow their own law, and, at a minimum, it is generally recognized that the rule of law requires the state to only punish citizens in accordance with the law – that is, to refrain from over-punishment (Gowder 2016, 7).

The problem with over-punishment from Caligula's amoral self-interested perspective is that it is widely believed to deter productive economic activity. If my property is not secure against the state – if there is a stated tax rate that is sufficiently low to permit me to profit from investment, but the real tax rate is substantially closer to 100 percent because of the risk of getting looted – then I'm much more likely to attempt to conceal my money or flee the country than to save or invest. And that means a smaller pie for Caligula to tax.

However, once again, there is a problem of short-term incentives to take into account. To see this, imagine again that Caligula is considering whether to punish an alleged tax evader, but, now, the person under her avaricious gaze is quite rich – and quite innocent of tax evasion. Nonetheless, Caligula is powerfully tempted to falsely accuse the rich person of tax evasion and steal all their goods, because, after all, it only costs $1,000 to do so, but the rich person has far more than $1,000 worth of stuff to steal. The time-inconsistency problem arises because Caligula's short-term and long-term interests conflict: If she could credibly commit to not punishing innocent rich people, she could give them an incentive to engage in productive activity, and hence collect more legitimate taxes in the long run. Thus, according to some scholars, the transition to the rule of law can be explained in part by the desire of leaders to maximize their long-run rents by building institutions permitting them to refrain from short-term expropriation.[28] In political science, Olson (1993; see also Haggard, MacIntyre, and Tiede 2008) argued that leaders of physical territory have an incentive to create functioning legal systems that restrain their own expropriative behavior in order to maximize the rents that they may gain from rule.

Dr. Strangelove's Tax-Evasion Punishment Machine could solve Caligula's over-punishment problem too. To do so, it must control the entire apparatus of punishment, it must only punish those who have failed to pay their ordinary taxes, and its functioning must be known and trusted by the public at large. But, how do we build it?

For platforms, the over-punishment problem is basically the same as the problem for states: In each case, the entity (platform/state) wishes to promote profitable activity (user engagement/capital investment) in the "space" it controls, but, in order to do so, it needs to provide some way to assure those whose activity is required that it won't just totally deprive them of the benefits of their own activity. What prospective influencer will build up a hundred thousand followers and a business based on their content if a slight shift in the political winds inside some company will cause that to all come crumbling down?

With respect to under-punishment, the platform problem is slightly different from the state problem. Platforms, unlike states, probably cannot usually inflict deterrent levels of punishment. At least with respect to social media platforms, with respect to most potential bad actors, the maximum punishment such a platform can inflict (a permanent ban from the platform) is almost certainly not going to be sufficiently painful to deter the worst misbehavior, such as by political propagandists, financial scammers, and the like (who may have teams of fake accounts and reliable ways to optimize on the cost of distributing their lies such that if their content or accounts are removed they will not have lost a too-large investment).

It may be that there is some lingering deterrent effect to the extent that if a platform is particularly effective at eliminating those who engage in rule violations, malicious actors may go looking for softer targets.[29] However, it will be safest to assume that the purpose of platform punishment is, in the classical typology of criminal justice,

[28] Another way to think about this is that Caligula's rate of discounting the future might change – if the regime seems unstable, it might be better to loot the citizenry now; if the regime is more stable it might be better to set up a system to protect long-run economic growth and get a smaller share of a much larger pie over a longer time.

[29] Another exception may be with respect to (a) businesses or politicians that are (b) heavily dependent on a given platform for their revenue or access to voters, and (c) have existing brand/political identities or other goodwill-type assets such that a platform ban is likely to be effective against strategies such as simply creating a new identity and rejoining the platform. Amazon and other transactional platforms might benefit from this kind of deterrent power; so might social media platforms when confronting famous influencers with distinctive individual identities such as real-life celebrities who heavily rely on social media.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

incapacitation rather than deterrence – at least in the social media context, platforms need to detect those who are creating a large proportion of the rule-violating content and remove them in order to control the distribution of that content.

Even though platforms lack the capacity to fully bring rule enforcement off the equilibrium path, they still have strong reasons to credibly commit to neutral rule enforcement in order to solve the under-punishment problem:

1. To the extent some marginal deterrent effect is possible, they can remove part of their enforcement costs from the equilibrium path.
2. Credibly committing to neutral rule enforcement may be able to keep some of the political pressure away from platforms. To the extent platforms can point to something akin to Dr. Strangelove's machine and say "See? We aren't making choices about what to enforce!" they are less subject to accusations of bias from both the left and the right.[30] More importantly, if company leaders are unable to succumb to political threats, for example, because someone else controls the rule enforcement system (or checks company control in a robust way), such threats cannot be effective.
3. There may be a marginal effect on user trust from credibly committing to enforcement – a company that makes credible promises to keep hate speech off its platform, for example, may have some competitive advantage, in terms of attracting new users and retaining existing ones, over companies that merely make unenforceable promises.

Because credible commitments solve the over-punishment problem and ameliorate the under-punishment problem, it behooves platform companies to figure out how to make them.

It is important to note that the two functions of credible commitment strategies – to bind an entity to a course of action, and to communicate that binding to external observers in a believable way – are distinct. Commitment strategies are important ways to enforce long-run–oriented behavior even independent of their capacity to signal credibility to outside parties. We might analogize a platform's ignoring or distorting its rules (banning a conservative for political reasons, failing to ban a harassing conservative) to individual health choices such as smoking or eating pizza. Smoking a single cigarette (ignoring a single powerful harasser) might produce more utility than its contribution to long-run pain; however, when this individual rational choice is repeated over an extended period of time, one ends up with lung cancer (an unsafe platform that drives users away).[31] Under such circumstances, some kind of precommitment strategy – that is, some way of making a long-term

[30] Thus, for example, Facebook's efforts to involve third-party fact checkers in its content moderation efforts in order to shield itself from accusations of bias (Lyons 2018).

[31] Cases like these are notoriously problematic from the standpoint of decision theory. I am inclined to see the problem as one of one-off decisions about *di minimis* risks which, when aggregated, are far from *di minimis*, however, this may be an incoherent way to see the problem (Lundgren and

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

decision to bind the organization to neutrally and completely enforcing its rules in the presence of short-run incentives to the contrary – is advisable.

However, in the platform context, because of the imperative platforms have to maintain user and public trust, a mere commitment to neutral rule enforcement, however ineffective, will be insufficient. Such commitment must in fact be known and believed by (i.e., credible to) outside parties.

4.3 ORGANIZATIONAL TOOLS FOR SELF-BINDING

For individuals, self-binding strategies typically involve recruiting the assistance of external coercion or technology – a set of techniques ranging from Odysseus tying himself to the mast to hear the song of the Sirens (Elster 2000) to applications that allow an individual to increase the cost of undesired behavior by, for example, setting up an automatic donation to one's least favorite politician to punish slip-ups.[32] Organizations, however, have more fine-grained control over their decision-making mechanisms, and can use institutional strategies to shape their own behavior. That is, they may change their own organizational structure in order to change the incentives shaping the entity as a whole.

4.3.1 *Independent Enforcers (Like the Meta Oversight Board?)*

One classic strategy that platforms may borrow from states is to change the identity of the person or entity who implements rule enforcement (including adjudication as a precondition of enforcement), in order to separate the actor who makes a decision about rule enforcement from the actor who feels the pain of the cost.[33]

I have suggested elsewhere that this strategy may play a role in the development of classic rule of law institutions in states, such as the independent judge (Gowder 2016, 59–62). Political leaders may create independent judges or other independent rule-enforcing institutions, and give them incentives to follow pre-existing law, as a precommitment mechanism to enable themselves to engage in costly rule enforcement (prevent under-enforcement). Doing so also protects against over-enforcement to the extent the independent judge doesn't personally receive the benefits of expropriation or is socialized to value legal compliance. Empowering a third-party, in other words, is how we get Dr. Strangelove's punishment machine.

Stefánsson 2020). Other ways of understanding such problems may be in terms of hyperbolic discounting or short-term failures of emotion regulation (Elster 2000, ch. 1). At any rate, the general pattern of such decisions will doubtless be familiar to readers.

[32] For example, Stikk, www.stickk.com/.

[33] Douek (2019, 24–26) draws on the literature on courts in authoritarian regimes to suggest that Meta's Oversight Board can help the company "outsource controversy" by providing a third-party to blame for unpopular decisions. This is an additional benefit of using independent enforcers for credible commitments, but not the most important one.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

This idea can apply to platforms as well. A key reason that top-level executives may be tempted to deviate from their own rules is because they are *personally* sensitive to the kinds of threats that might be posed to the company as a whole. Mark Zuckerberg may have been particularly sensitive to the threats of Republicans, causing him to under-enforce Facebook rules against right-wing rule-violating users, because Zuckerberg personally loses a lot of money and status if Josh Hawley successfully takes Facebook's Section 230 exemption away. His employees would be worse off too, but, because, in economic terms, their wealth is vastly more diversified (it's mostly in kinds of human capital that they can convert to cash by working for other companies), they have much less of a felt need to surrender to potentially empty threats. This partly explains why line employees have tended to be stronger advocates for rule enforcement, according to the media accounts cited above, than senior executives: Intuitively, more senior executives are likely to have more firm-specific capital (be less diversified), for example, by having their reputations, networks, and knowledge tied to a specific firm (and perhaps also a higher degree of investment in the firm's stock, as well as compensation packages more tied to the firm's performance).[34]

Unfortunately, judicial independence or anything analogous to it tends to be difficult to achieve and sustain in the world of states because it conflicts with some fundamental imperatives of leadership: Top-level leaders have strong reasons to centralize power in order to maintain their leadership and policy autonomy, and handing over authority to independent enforcers, along with enough sources of power (money, military force) to enforce their own independence undermines that centralization. Thus, trying to use independent enforcers to help leaders constrain themselves may just push the problem back a step: Instead of struggling to commit to costly enforcement of their rules, leaders now struggle to commit to maintaining the independence of their enforcers. This too is a problem for platforms; witness the skepticism about the genuine independence of Meta's Oversight Board in view of the company's control over things like the information it receives and the selection of its initial members (e.g., Newton 2022).

Platform enforcer independence may be easier to achieve than state enforcer independence, if only because platform enforcement does not require the direct application of physical coercion. In the physical world, independent enforcement has to be created by law and backstopped by force, but in the platform world, it can be created by, as Joel Reidenberg (1998) and Larry Lessig (1999) taught us over two decades ago, code. Some degree of platform enforcer independence could be achieved as a purely technical matter, by, that is, engineering direct control over

[34] By way of caveat: The capacity of workers to constrain the companies they work for is limited not only by the relative balance of interests and economic power but also by the ideologies that firms and workers develop to justify what they do. Ari Waldman (2021) illustrates this best in an insightful study of how the concept of "privacy" becomes warped within the workplaces of companies organized around its opposite.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

such decisions to personnel within the authority of the enforcer. For example, an independent enforcer for Twitter could have software-level control over the decision to ban or not ban purveyors of disinformation, with the software in question being subject to a third-party audit to ensure that no "back doors" were available to override those decisions.

However, many of the same problems that vex states may still arise in different forms in the context of platform enforcement. There is an inherent trade-off between organizational policy autonomy and the existence of independent enforcers: An independent enforcer, in virtue of its independence, has the capacity to defect from centralized policy decisions. For example, commentators have speculated that Meta's Oversight Board could effectively set aside the company's policy on political advertisements (Levy 2020b).

Moreover, independent enforcers have their own organizational capacity or lack thereof, which may affect the policy/independent enforcement trade-off. Consider, for example, the problem of caseload management: An independent judge who has the capacity to hear many cases (a large staff budget, an efficient adjudication process) can systematically distort policy by defecting from it; an independent judge who only has the capacity to hear a few cases cannot effectively ensure that pre-existing rules are enforced. Hence, a leader trying to empower an independent enforcer still has to make difficult choices even given the ability to use code to entrench its power: Give that enforcer too little organizational capacity, and it may not be able to sufficiently support the kind of credible commitment that the organization needs to make to the outside world; give it too much organizational capacity and it might start imposing its own preferred policies on the broader entity. Put differently, independent enforcement that is effective tends to also entail the delegation of policy autonomy, and there are significant challenges in delegating that policy autonomy in a legitimate fashion; this is, I submit, the foundation of many conventional challenges to constitutional judicial review in modern polities: If we give judges enough institutional capacity to effectively enforce the constitution, we also risk giving them enough institutional capacity to illegitimately impose their own policy choices on elected leaders.

4.3.2 *The Political Foundations of Credible Commitment: Recruiting Workers and Ordinary People to Backstop Self-binding*

In the context of states, many self-control problems are mitigated by democratic institutions that permit mass publics both to exercise some control over policy and to backstop (i.e., by their capacity to sanction political leaders) the independence of judges and other enforcers (Gowder 2014b; Law 2009). This allows for the incentives of policymakers and enforcers to be sufficiently aligned to reduce the risks of enforcement defection, in virtue of the fact that the power of each depends on the willingness of a mass public to support their decisions. For that reason, it is less risky to confer additional organizational capacity on enforcers.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

In order to understand the underlying strategic dynamics, please indulge me in a brief digression on how it works for states. Democratic publics suffer from two principal-agent problems: Once they've put an executive in office, that executive suddenly commands a lot of force and engages in a lot of hard-to-observe behavior; how to keep him or her to the will of the public and the laws that (ideally) represent that will in the long run? But, on the other hand, how to keep judges from excessively impeding the pursuit of policy goals by the executive that the people support? The solution is for the judges to control executives, but for this control to go through the threat of collective action by the people, which will only occur when the judges don't defect too badly from the people's present will. That is, in a reasonably well-organized state, the decisions of independent judges who are institutionally committed and socialized to value legal propriety can be used as a signal to trigger coordinated public action (e.g., voting the recalcitrant executive out). So long as those decisions are reasonably well-aligned with the preferences of the public at large, and so long as the public has the capacity to observe executive defiance of judicial officials and engage in collective action, if executives disobey, the public can coordinate on disobedience as a signal to inflict political punishment on executives.[35]

Potentially, platform companies could make use of similar mass-directed policy and enforcement alignment. The "mass" in question could be either (or both) of their employees or their userbases (or even the general public, with some caution about defining that public in an international context and its relationship to a userbase). I will take them up in turn.

Prominent cases of employee activism at many major platform companies suggests that employees have some capacity for collective action – at least during time periods when the technology industry is flush with money for workers (in times of layoffs and contraction, presumably worker power decreases).[36] For example, Google employees organized to prevent the company from doing ethically dubious work for the Pentagon (Wakabayashi and Shane 2018), and Microsoft and Amazon employees extracted at least token concessions from corporate leaders about climate change and work for Immigration and Customs Enforcement (ICE) (Gurley 2019; Frenkel 2018).[37]

Some scholars have suggested that the constraint of state power in historical states has at least in part arisen from the existence of such alternate sources of power within an organization. In premodern states, independent holders of "administrative power" (such as feudal lords), in governments with limited capacity to centralize, have

[35] Obviously, this solution is imperfect, as evidenced by the sorry state of the U.S. Supreme Court right now, which illustrates the difficulty in circumstances of extreme polarization among the public of preventing that polarization from infecting a court; for more details, see Balkin's (2020) account of constitutional rot.

[36] For a discussion of several examples, see Srivastava (2021, 8).

[37] At a smaller software company, a single employee apparently sparked the cancellation of an ICE contract by deleting code that he had written which was being used by the agency (Cox 2019).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

extracted concessions from rulers that have led to the constraint of top-level leaders (De Lara, Greif, and Jha 2008). Organized employees may serve as something analogous to the holders of administrative power, to the extent that there is some policy sweet spot that is consistent both with the long-run interests of the company as well as with employee values and interests. If there is such a sweet spot, then employees can demand that corporate leaders follow independent enforcer rulings to the extent those rulings are faithful to that "sweet spot" policy, and hence work together with the independent enforcer to allow corporate leaders to commit to a more long-term oriented rule enforcement strategy. In effect, this becomes a perspective shift on the right-wing critique noted at the beginning of this chapter: Maybe rather than creating short-term defections from company policy, employees can prevent them.

The "administrative power" approach also appears in the literature on contemporary authoritarian governments.[38] A company is a kind of benign authoritarianism as to its internal operations, and some companies are more authoritarian than others, depending on matters such as the extent to which a company is insulated from market controls on its leadership – the example that immediately leaps to mind is Meta, insofar as Mark Zuckerberg's famous stock ownership arrangement (Durkee 2019) effectively guarantees that he can be CEO as long as he wants.

Anne Meng (2020) recently published a monograph beginning with the puzzle of how some authoritarian leaders can transition from a personal regime that cannot survive the death of the leader to a stable authoritarianism (like contemporary China) – a transition that hinges on the building of institutions that stand apart from top-level leaders and can constrain them. But, at a sufficient level of abstraction, this is just the question that we're presented with corporate platforms, and so the same kinds of insights into how it was possible for the Tanzanian National Assembly to constrain Julius Nyerere could shed light on how we might make it possible for the Oversight Board to constrain Mark Zuckerberg.

Meng's answer is that successful institutions empower durable elites other than top-level leaders, which then allows other elites to effectively make alliances (i.e., solve a collective action problem among themselves) sufficiently strong to counteract top-level leaders. In her words:

> When an elite is given a key cabinet position, such as vice president or the minister of defense, he is given access to power and resources that allows him to consolidate his own base of support. Elites who are appointed to positions of authority within the regime then become focal points for other elites. They become obvious potential challengers to the incumbent if she were to renege on promises to distribute rent.[39]

[38] Cf. Douek (2019), who also borrows from the authoritarianism literature.

[39] Meng (2020, 16). Note that Meng isn't developing a theory of third-party enforcement as such. She focuses on the need for authoritarians to build institutions in order to preserve their own rule – in effect, an authoritarian who does so is recruiting allies by, for example, putting someone else in charge

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

Platforms might also deploy such "elite" kinds of constraint, such as by empowering executives in ethics or oversight roles to constrain those in product and marketing roles who might otherwise undermine platform governance efforts.

Relying on more traditional workplace power management tools, a unionized workplace could write control over rule enforcement into a labor contract, with that control to be enforced by a combination of legal and labor action, or governments could impose regulations providing for intraorganizational insulation of decision-making functions, analogous to regulatory strategies currently used within the financial industry which require separation of functions and of information within organizations and their contractors. For example, the US Sarbanes-Oxley Act prohibits external auditing firms from offering some nonaudit services to the companies that they audit, and the European Union regulates the amount of revenue that auditors may receive from nonaudit services (see Gelter and Gurrea-Martinez 2020, 808–11 for references). Similarly, regulations in many countries require internal boundaries between employees who participate in trading functions and employees who have access to insider information, or effectively require such policies by using their existence as a factor in decisions about insider trading enforcement actions (Dolgoplov 2008; Dahan et al. 2012, 222).

Such an internal separation of powers strategy is readily available to companies and their regulators, for example by requiring the separation of policymaking and enforcing functions from lobbying functions to keep those personnel most susceptible to political pressure away from governance. The same is true about the merger between rule enforcement and customer service noted above: Companies or their regulators may enforce a separation between account managers that service big advertisers or relationship personnel servicing major users and rule-enforcement functions.

A variety of other institutional designs might be available that involve integrating employee decision-making capacity with some other actor's decision-making capacity in order to tune the degree to which a company tracks short-term rather than long-term interests. At the limit, such a strategy amounts to constitutionalizing the operations of a company via workplace democracy.[40] In short, there are many ways that existing company employees, whether junior or senior, could be empowered

of the military or (particularly importantly for Ming) creating a legally designated successor, who then becomes invested in the stability of the overall regime so long as the top-level leader continues providing benefits. By contrast, this chapter focuses on the need for leaders to constrain themselves to make long-term commitments, which they want to do in order to generate stable expectations in others (i.e., trust), which will in turn allow them to draw more benefits – whether that's rents from rule, qua Mancur Olson, or stock value from owning a platform company. At the most abstract level, however, this more or less amounts to the same idea, viz., that top-level leaders can empower and recruit lower level elites to backstop their ability to credibly make promises in the context of shared benefits and cooperation.

[40] On constitutionalizing platforms, generally, see Suzor (2019); on workplace democracy, see Landemore and Ferreras (2016).

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

either to backstop an independent enforcer or to serve independent enforcer functions (or other leadership constraint functions) themselves.

However, because technology industry workers do lean to the political left on at least some important issues, institutional innovations that deploy organized employees to prevent short-term corporate failures of self-control may be more effective with respect to the critique from the left than that from the right – that is, handing greater control over content moderation to employees may bring it about that Donald Trump gets banned for threatening violence in violation of Twitter policies, but may not do a lot to prevent conservatives from getting shadowbanned.

4.3.3 *User-Generated Sanctions for Company Commitment*

What about the users and the rest of us? In international relations, one prominent strategy for achieving credible commitment is creating *audience costs*: making leaders vulnerable to sanctions from the general public if they violate their commitments. Fearon (1994; see also Tomz 2007), for example, argued that democratic states have the capacity to buttress the credibility of their public escalations in international military crises, in view of the fact that the leaders of democracies are accountable to domestic audiences which may impose sanctions on them for backing down after vigorous saber-rattling. The "audience" in the theory supposes a third-party to the transaction: The leader of one country threatens the leader of another, with the threat made credible by the external sanctions posed by the first country's citizens. But the idea of making oneself subject to sanctions by some outside party in order to backstop a commitment is more general, and can be applied just as well when the party applying the sanctions overlaps with the party to whom the commitment is directed.

If a sufficiently large group of users to inflict short-term pain on a platform company has the capacity to act collectively, then the company has a short-term incentive to keep from offending them. If companies (or the rest of us) can exercise some influence over the extent to which their users have the capacity to act collectively, to monitor their behavior, and otherwise to effectively inflict sanctions, then the potential exists for institutional design to affect whether platforms can effectively carry out their long-term interest in neutral rule enforcement.[41] This may be particularly appropriate as a strategy under circumstances in which company employees are not trusted, such as with respect to the right-wing critique of social media.

I propose to reinterpret the many existing calls for greater transparency (e.g., Suzor 2019, 136–41; Gillespie 2018a, 198–99; Suzor et al. 2019) in social media content moderation as incomplete suggestions along these lines. Arguments for transparency as a primary solution to the challenges of content moderation make sense if we suppose that external constituencies have some degree of latent power to sanction

[41] Cf. Gowder (2018b), suggesting technological approaches to coordinated consumerism in other contexts.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

platform companies, and merely lack the information to exercise it appropriately.[42] However, transparency solutions cannot actually work if the problem is not a lack of information but the inability to engage in coordinated action to deploy sanctions.

Unfortunately, the status quo makes coordination particularly difficult for platform users. The strong positive network externalities of dominant platform membership mean that it takes a much larger group of users than it otherwise might to credibly threaten to punish a platform. In less jargony terms: The only way users have to sanction platforms right now is by foot-voting (disengaging from, or quitting, platforms). And perhaps people might want to quit a platform in response to its non-neutral rule enforcement, but they might nonetheless benefit too much from being on the platform to do so, unless they can get lots and lots of other people to go with them. Witness the difficulty that many journalists, academics, influencers, and the like are currently experiencing as of this writing (December 2022) in leaving Twitter after the Musk acquisition.[43]

While this effect might undermine the incentive for platform companies to observe their own rules in the short run, it may harm their interests in the long run. A sustained pattern of inconsistent rule enforcement might eventually reach a tipping point at which a company can no longer retain the loyalty of users or its capacity to recruit new ones, at which point the entire ecosystem comes crumbling down and the platform experiences a sudden (but difficult to foresee) mass abandonment.

The nightmare scenario for a platform company would be a kind of abandonment cascade structurally similar to the preference falsification cascades Timur Kuran (1991, 1989) has analyzed. Suppose that different subgroups of users have different levels of tolerance for inconsistent rule enforcement (censorship, failure to get rid of hate speech), where those levels of tolerance also are increasing in the number of users on the platform due to network externalities. Then an abandonment cascade could occur if a platform acts so inconsistently (or appears to do so) as to drive away group A, which (because the degree of positive network externality for

[42] Transparency solutions are challenging for platforms because of the tension between public rules and operational security: Fully public criteria for user behavior and methods for controlling it are unlikely to be sustainable in an environment where sophisticated organized actors (i.e., Russian intelligence agents, among others) are dedicated to subverting platform mechanisms for malicious purposes. This challenge is significant for states as well, of course, as is represented by perennial debates in American law between the paradigms of criminal justice and national security in the war on terror. But at least states have the advantage of a relatively clear-ish distinction between domestic and foreign actors, and secure versus insecure spaces; whereas for platforms everyone in the world is a "citizen" – Russians are perfectly legitimate users of Facebook, and are just as entitled to occupy the same discursive spaces as everyone else, so long as they aren't trying to subvert other societies on it.

[43] In this context, it is at least suggestive that the most prominent effort to quit the major social media platforms has been from the so-called "alt right," a collection of political extremists with a substantial existing alternative media ecosystem that probably made it relatively less costly for them to coordinate to switch from Twitter to Gab, Parler, Truth Social, and so forth – especially since many of their most prominent members had already been chased off the mainstream platform, and hence their presence could not provide an incentive for others to stay.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

everyone else shrinks when A is gone) lowers group B's toleration, and hence drives them off; with B gone, C's toleration decreases, and so forth. Such a cascade, as with Kuran's revolutions, could come very suddenly.

There is some evidence to suggest that the cascade model accurately captures platform user incentives: Scholars have identified something like an abandonment cascade in the collapse of Friendster (Garcia, Mavrodiev, and Schweitzer 2013). If this is right, then users may have no effective capacity to collectively (and purposively) threaten sanctions against platforms, but nonetheless pose the real prospect of totally destroying a platform with their aggregate, emergent behavior, out of nowhere, in an abandonment cascade. Platforms might be able to stave off this risk by empowering users to sanction them for trust-betraying behavior before that behavior reaches the point where an abandonment cascade happens – by credibly threatening or inflicting a punishment short of leaving the platform. But merely providing information to users won't do the trick – there must be some noncascade and precascade way of using that information.

This suggests that calls for transparency, at least in the absence of a plausible account of how the general public might actually exercise leverage over companies, are insufficient as a strategy for constraining companies. To be sure, companies have other reasons to operate more transparently in their rule-enforcement. If they discover an effective method of controlling company behavior, transparency about that method is imperative in order to ensure that external stakeholders can actually observe that a company's commitments are, in fact, credible. But the core problem is the structure of the sanctions that might be imposed on a company in the context of the network externalities described above: Platforms are unlikely to have sufficient incentive to keep their users pleased, transparency or no transparency, until the point where an abandonment cascade hits, at which point it's too late. In order to change the shape of this sanction curve, it becomes necessary to provide more intermediate levels of sanction by users or the public at large.

Another way to think about the call for transparency and the broader problem of user-generated sanctions, which might help point the way to more effective techniques, is as an argument for changing the nature of "insider" and "outsider" status with respect to platform conduct. In order for outside actors to constrain companies, some of the knowledge – *but also some of the control* – that had previously differentiated insider versus outsider status – which content is deleted, how the decisions are made – will need to change character. And the company resistance to transparency is partly explained by the need to maintain that status differentiation: By keeping leverage over rule enforcement to insiders, who are vetted by hiring processes and kept loyal by paychecks, platforms ensure that the interests of those who have access to a source of power over their operations are aligned with their interests in areas such as maintaining revenues and protecting against security threats.

But a wide variety of intermediate statuses between full insider and full outsider are possible. Meta's Content Moderation Oversight Board is one example: By recruiting

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

carefully vetted outsiders who have divergent interests from the company, but whose interests are (hopefully) aligned with the general public rather than with, for example, Russian attackers, and giving them a degree of privileged, insider-level, access to information and authority, Meta does not just propose to give [quasi-]outsiders *knowledge* about what the company is doing, but also *power* over it. If the Oversight Board is to work in the long run, its members must be able to either directly control company behavior (e.g., via technical means), inflict or provoke sanctions on company leaders, or increase the salience and credibility of the existential risk of abandonment cascades if their decisions are defied (and hence make abandonment cascades less likely by giving company leaders a clear message about what behavior is necessary to avoid them).

Another kind of technique to mediate between insider and outsider status has appeared in the corporate law context: the rise of the benefit corporation (Hiller and Shackelford 2018). By making business leaders formally accountable to stakeholders other than their shareholders, such forms of corporate organization have the potential to give those stakeholders some power to exercise constraint over the companies (although corporate law scholars of my acquaintance tell me that this has not been realized in practice) – or, we might say, to radically expand the group of people who are considered "owners," and hence have authority over the organization. More generally, forms of corporate organization like the benefit corporation or other ideas from the broad field of "stakeholder capitalism" (Freeman, Martin, and Parmar 2007; Gadinis and Miazad 2021) might be deployed in order to give outsiders nonexistential threats over platform companies. For example, such tools might be used to backstop neutral platform rule enforcement by giving organized groups of outsiders the legal power to enforce neutrality by filing lawsuits for money damages, the way shareholders in ordinary pure for-profit corporations can (theoretically) file lawsuits against corporate executives who are insufficiently attentive to their duty to maximize profits. Company leaders may have an incentive to confer the capacity for such suits on outsiders as a commitment strategy to publicly force themselves to take the actions most consistent with a company's long-run success.

Here is also where capacity-building efforts by governments may make a particular difference. The rules of contract and of corporate law are controlled by our democratically elected governments, and are malleable. We should consider using them to give platform companies – and others – the tools to subject themselves, in a controlled fashion, to some kind of authority beyond the whims of their leaders. For example, we can give legal teeth to the notion of a benefit corporation, and by doing so confer on companies the tools to in turn confer on the general public some capacity to use the courts to enforce their compliance with public-oriented missions such as providing neutral platforms for speech and sociality. We can create forms of stock ownership that can be conferred on public interest groups to give them some degree of direct leverage over companies. And we can modify labor law to provide

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

employees with the tools and the incentives to exercise greater voice within companies in order to backstop leaders' commitment to their own public purposes. Some preliminary sketches along at least the last of those lines appear in the conclusion to this book, although as a whole I leave this particular dimension of the approach to corporate governance scholars for further development.

4.4 TOWARD PLATFORM RULE OF LAW

There is a widespread sense that platforms have begun to exercise government-like power, but without the kinds of constraints, such as democracy and the rule of law, that keep government power in check. This idea has been increasingly popular in academia, with, for example, scholars such as Nicolas Suzor (2019) arguing for a new constitutional settlement to bring this power under control, and Rory Van Loo (2021) arguing for extensive procedural protections in their adjudicative processes. Platforms allegedly exercise quasi-governmental power not only in content moderation, but also in copyright enforcement, particularly in the American DMCA "notice and takedown" regime (e.g., Perel and Elkin-Koren 2016), and in business marketplace regulation and dispute resolution, particularly among transactional platforms like Amazon (Van Loo 2016). This sense has also evidently leaked out into the public at large, as evidenced by (thus far unsuccessful) attempts to extend US First Amendment protections (and their state law equivalents) to platforms via formerly obscure doctrines previously extended, at the federal level, only to company towns,[44] and, at the state level, primarily to shopping centers and similar "functional equivalents of the traditional public square."[45] Thus far, such lawsuits have universally failed, but they represent a strong indication that at least some among the public take seriously the notion that platforms are illegitimately exercising quasi-governmental powers.[46] The same is true of legislation enacted in Texas and Florida as of this writing, which purport to prohibit social media political "censorship."[47]

The existence of such legitimacy challenges also creates a compliance challenge. There is well-known empirical research suggesting that compliance with the law depends in part on perceived fairness (Tyler 1990, 1997). From that research, we can predict that to the extent platform rule enforcement is perceived as inconsistent and unfair, users will be less willing to obey those rules.

Accordingly, it will be advantageous for platforms to develop a kind of internal rule of law. By this, I mean systems of constraining their uses of power which follow,

[44] *Marsh v. Alabama*, 326 U.S. 501 (1946).

[45] *Robins v. Pruneyard Shopping Center*, 23 Cal.3d 899 (1979).

[46] See, for example, *Prager University v. Google*, 951 F.3d 991 (9th Cir., 2020) (rejecting First Amendment suit against Google for putting a right-wing nonprofit's YouTube videos in "restricted mode").

[47] As of this writing, both laws are actively winding their way through constitutional challenges with the US Supreme Court as their inevitable destination.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

more-or-less loosely, three principles that I have articulated in the context of states (Gowder 2016), adapted for the platform context:

1. Regularity: Platforms should follow their own rules; they should refrain from sanctions or other exercises of authority against users for the platform equivalent of "reasons of state," that is, short-term profit motives.
2. Publicity: Users of platforms should have the opportunity to know the rules that apply to them and to contest their application to their conduct in a procedurally fair way.
3. Generality: Platform rules should be created and applied in a way that recognizes the equal standing of all platform users, regardless of, for example, nationality, gender, race, religion, or political orientation. While this principle does not forbid the differential treatment of some categories of users – for example, the banning of neo-Nazis from social media – it does require that such differential treatment be publicly justifiable in terms that recognize the equality of all to whom the rules are addressed (a criterion that Nazi bans manifestly satisfy).

The implementation of organizational changes to facilitate such a platform rule of law – the creation and support of independent enforcers, and the strategic integration of employee power and outsider scrutiny into rule creation and enforcement processes – has been the topic of this chapter. If legitimacy and compliance scholars such as Tom Tyler are right, such organizational changes may also help solve platforms' broader governance problems by facilitating user compliance.

There is also a moral reason for such an endeavor. We have normative requirements for legal and constitutional institutions, and those requirements may be sensibly applied in part to platforms, generating questions such as:

- Are decision-making institutions really and truly independent of those with interests in the decisions?
- Are there determinate rules that actually bind the platforms (their leaders)?
- Were those rules made in a way accountable to those who are supposed to benefit from the rules?
- Are decisions made in a fair way?

The widespread answer among the public to most of those questions is "no." But these are precisely the issues that the international rule of law development enterprise has concerned itself with (at varying degrees of competence and avoidance of colonialism) in the world of states. In my own work on the rule of law, which attempts to learn from those experiences and from the history of the rule of law, one of the key conclusions that I've drawn is that effective rule of law institutions almost always depend on the threat of collective action by sub-elites and/or the great mass of people to hold the powerful to account.

As applied to the platform context, this suggests that rule enforcement institutions, and the rules themselves, have to be sociologically legitimate. If the Oversight

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press

Board says "Meta has to do X," and we want to deploy market sanctions in order to force Meta to actually do X, then the decisions of the board have to be legitimate enough to motivate people to get mad and seek out alternative platforms if Meta disobeys. It also requires that people have some way of engaging in collective action. If they're actually motivated by forcing a company to stick to its own rules, they need effective tools to act on that motivation, that is, to learn from trusted rule-interpreters when the company has broken the rules, and to coordinate their behavior.

This is, ultimately, the recipe for the credible commitment of a powerful entity to follow its own rules: public, decentralized, power that can be collectively used to hold rule-enforcers to compliance, backstopped by credible and legitimate monitors administering credible and legitimate systems of rules and signaling to the public when the rules have been broken. And this entails a deep integration of ordinary people in governance. For the rules themselves need to be the sorts of things that people want to collectively enforce. The obvious path to this is to make them democratic, that is, to give the people, whether citizens, users, democratic governments, civil society, and so forth, some say in platform rules, rather than having it just be Mark Zuckerberg who determines the conditions under which someone like Donald Trump can or cannot post insurrectionary material on social media. The rule of law development framework can at least give us some criteria for telling whether we've succeeded: We can say that a rulemaking process is democratic in the right way if it forces the kind of alignment between people and interests that can actually draw on collective action for its support. Such democratic institutions can help achieve the platform rule of law insofar as participatory rulemaking has a legitimating function *and* participatory adjudication provides ordinary people with information about one another's interests and beliefs about platform conduct in order to facilitate coordination.

Chapter 6 of this book sketches a preliminary design for some of the institutions that might be put to work to bring this about. But before getting there, we should look at the most developed existing attempt to create an independent enforcer for a platform rule of law to test it against the theoretical material developed thus far. That is the task of Chapter 5.

https://doi.org/10.1017/9781108975438.005 Published online by Cambridge University Press