5

Actually Existing Platform Self-constraint … Up to a Point

The Meta Oversight Board

As of this writing, the most assertive step toward building an institution potentially capable of meaningfully expanding the capacity for platform governance is Meta's Content Moderation Oversight Board.[1] Conceived in 2018 (Klonick 2020, 2449–50) and consciously modeled on something like the US Supreme Court, the Board issued its first decisions on January 28, 2021.[2]

This chapter considers the model of platform governance which the Board represents in the context of the problems raised by the rest of the book. It is in part a qualified defense of the Board: I argue that an entity like the board can help platforms build short-term responses to emergencies like the January 6, 2021 autogolpe attempt into sustainable long-term rules. However, ultimately, no "Supreme Court"-like entity can solve the problems considered in the previous chapters on its own. Rather, platform adjudicators ought to look less like judges and more like juries – for the knowledge they are required to deploy is not specialized expert knowledge on rules of law but contextual grounded knowledge of the conditions of their local environments and the interaction between platform activities and those local contexts.

[1] The primary sources for the description of the Oversight Board and its purposes and history in this chapter are the following: (1) The charter of the Oversight Board, as posted online at https://scontent-ort2-2.xx.fbcdn.net/v/t39.8562-6/93876939_220059982635652_1245737255406927872_n.pdf as of May 26, 2020 (cited in this chapter as Charter, by section and subsection); (2) the Bylaws and Code of Conduct of the Oversight Board, as posted online at https://scontent-ort2-2.xx.fbcdn.net/v/t39.8562-6/93836051_660280367850128_4544191419119566848_n.pdf as of May 26, 2020 (cited in this chapter as Bylaws or Code of Conduct, by article and section); (3) the June 27, 2019 public consultation report released by Facebook, as posted online at https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-2.pdf (report) and https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf (appendices) as of May 26, 2020. Note that Appendix E to the Consultation Report is the comparative institutions report co-authored by me and Facebook's Director of Product Policy Research, referenced above, for which I was paid – see the appendix to the introduction for details; and (4) an op-ed by the four co-chairs of the Oversight Board, Catalina Botero-Marino, Jamal Greene, Michael W. McConnell, and Helle Thorning-Schmidt, "We Are a New Board Overseeing Facebook. Here's What We'll Decide," *New York Times*, May 6, 2020, www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html.

[2] Oversight Board decisions, www.oversightboard.com/decision/.

https://doi.org/10.1017/9781108975438.006 Published online by Cambridge University Press

The defense of the Oversight Board noted above also entails developing some more conceptual and normative ideas about the notion of platform identity. This chapter sketches a kind of platform legal identity similar to the constitutional patriotism developed by some scholars in the context of states, about which I have written elsewhere (Gowder 2019). I contend that this theoretical work might contribute to addressing some of the underlying controversies associated with the power of platform companies.

5.1 WHAT FUNCTIONS MIGHT THE OVERSIGHT BOARD SERVE? DOES IT DO SO WELL?

In the abstract, we might categorize the functions that an Oversight Board might carry out for Meta and for the outside world into six buckets: (a) propagandistic, (b) informational, (c) corrective, (d) constraining, (e) reformist, and (f) inclusive. I shall ultimately argue that there is a seventh function, which we can call "rationalizing," which the Oversight Board is most likely to serve – and which fits into a broader story about a kind of platform rule of law. But that last one requires rather more theoretical development, whereas the original six buckets are somewhat more conventional.

Propagandistic functions include insulating Meta from external criticism by creating the appearance of oversight and encouraging the perception that decisions are attributable to neutral outsiders. This description is self-consciously neutral as to whether that perception matches reality or not.

While "propaganda" carries negative connotations, not all propagandistic functions are necessarily bad; for example, to the extent Meta is subject to political threats on the basis of false claims of partisan bias in content moderation and to the extent those threats depend for their political force on convincing the public that company personnel are deliberately engaging in political censorship, the propagandistic effect of the Oversight Board may be beneficial for the rule of law-esque reasons discussed in Chapter 4. If the Board is also trusted to make fair decisions, it could improve adherence to those decisions (i.e., reduce efforts to evade content policies or protest and resistance to them) and shield Meta from external political pressures by promoting their broad-based sociological legitimacy. Thus, Klonick (2020, 2426) observes that the Board may be a "convenient scapegoat for controversial content-moderation decisions." But this might be a good thing: If the Board is a scapegoat for *rule-compliant* content-moderation decisions, then that amounts to insulating company executives from paying the political costs of their compliance, and hence facilitating the credible commitments described in Chapter 4.

However, there is also a dark side to the notion of "propaganda" to the extent the Board also insulates Meta or its executives from the pressure they *ought* to experience – more generally, the Board may also amount to what Flew and Gillette (2021, 240) characterize as "pre-emptive self-regulation" which "inhibits the development

of a regulatory framework for platforms as a whole." The capacity of the Board to serve propagandistic functions for good or for ill largely depends on how credible its independence and authority are to outside observers.

The notion of an *informational* function identifies that adjudicators can surface information about governance problems by giving individuals who experience those problems an incentive (in the form of the increased likelihood of having their complaints satisfied) to communicate that information to the public, regulators, and the company itself – with communication to the first two of those mediated by Board decisions (which are public). That potential is present for a wide variety of problems within and without the company – an Oversight Board case could draw attention to some way in which otherwise-reasonable platform rules were causing unintended harm due to novel external circumstances, or to the way in which platform rules and processes themselves are unreasonable on their own terms or underenforced. For an example of how this is already occurring, note that the Oversight Board's adjudications, in conjunction with media reports, have recently drawn attention to troubling features of Meta's internal governance such as the "cross-check" system (discussed in Chapter 4). In doing so, the Board potentially subjects the company to greater public accountability (Schechner 2021).

A *corrective* function is simply the capacity to fix individual incorrect decisions, relative to some standard that includes – but is not necessarily limited to – consistency with platform rules. In view of the relatively low stakes of most individual decisions, this is in some sense the least interesting function of any adjudicator. But sometimes the stakes are high, with the quintessential example being the Donald Trump ban which this chapter considers in detail.

Constraining functions were the subject of Chapter 4. In the context of the present typology, we can understand constraining functions to simply be the aggregation of large-scale informational and corrective functions: That is, by identifying deviations from pre-existing company commitments (in the form of content moderation rules) to internal and external constituents with the capacity to sanction decision makers, and by identifying the commitment-complying (correct application of the rules) decision, an adjudicator can give those constituents the resources to effectively demand that decision makers follow their prior commitments (Hadfield and Weingast 2013). This, of course, depends on its genuine independence for the reasons described earlier.

Reformist functions are closest to those of a stereotypical activist constitutional court, such as the Warren Court in the United States. An adjudicator with sufficient capacity to enforce precedential decisions can directly modify the policies of those whom it regulates by decreeing new rules.[3]

[3] That capacity might come from the ability to decide cases in bulk and hence directly implement those decisions, or from sufficient legitimacy to motivate empowered third parties to enforce those decisions in new contexts.

https://doi.org/10.1017/9781108975438.006 Published online by Cambridge University Press

Finally, *inclusive* functions capture the overarching democratic aims of this book. An adjudicative body can have a set of decision makers different in morally or practically important senses from those responsible for the underlying decisions which it reviews. Accordingly, it can supply otherwise neglected constituents with an avenue to influence outcomes with respect to any of the other functions. In the context of Chapter 3 of this book, inclusivity can also mean *localism*, that is, incorporating knowledge from those who are closer to the site of some governance challenge.

To some extent, the Board makes improvements in these respects at least from the baseline state of affairs at Facebook/Meta before its creation. The requirement that board panels have a representative "from the region which the content primarily affects" introduces some degree of localism and inclusivity to its decisions.[4] However, in view of the fact that the "regions" are extremely large, in some cases entire continents or bigger, and of course that the board is an elite institution, it is unlikely that significant local knowledge could be incorporated in this fashion.[5]

The financial arrangements for independence (key to the constraining function) are also – at least tentatively – convincing. So far Meta has contributed almost three hundred million dollars to a trust to support the Board's arrangements, giving some reason to believe that (assuming that it doesn't retain control over trust personnel or decisions) the Board will be capable of being reasonably independent at least until the trust money runs out. However, as is always the case with adjudicators operating with limited time horizons, there is reason to worry that Board members may be tempted to shape their decisions to curry favor for future employment opportunities. The fact that Board decisions are unsigned may mitigate this risk.

However, the Board will face several key challenges. First is institutional capacity. The fundamental struggle for all platform content moderation efforts is the sheer volume of cases to be considered, and the Board's likely inability to hear a truly large quantity of cases without (for reasons described below) undermining its independence creates an upper limit on its capacity to review company decisions.

If the Board can only hear a handful of cases relating to particularly salient or policy-relevant (i.e., precedential) conflicts, then it may be able to provide some external

[4] Board Bylaws Art. 1, Sec. 3.1.3; for discussion, see Klonick (2020, 2471). Additional Meta localism exists in its "trusted partner program" by which it seeks input from civil society organizations around the world, however, it is unclear how much actual influence such organizations have. See Meta Transparency Center, "Bringing Local Context to our Global Standards," January 28, 2022 (updated), https://transparency.fb.com/policies/improving/bringing-local-context. Moreover, it's unclear whether the purpose of this program is genuinely to seek input on content rules or simply to comply with hate speech regulations, as an outcome of the negotiation between several platform companies and the European Union described by Bloch-Wehba (2019, 45).

[5] The regions are: "United States and Canada; Latin America and the Caribbean; Europe; Sub-Saharan Africa; Middle East and North Africa; Central and South Asia; and Asia Pacific and Oceania" (Bylaws Art. 1, Sec. 1.4.1). As Douek (2019, 33) points out, a member from one part of a particularly diverse region – such as "Sub-Saharan Africa" or "Asia-Pacific and Oceania" – is unlikely to be all that capable of applying local knowledge to another part.

input into difficult policy decisions, but it is unlikely to be able to control enough outcomes or hear enough complaints to exercise a constraining role or support broad-based inclusion or procedural justice and hence its legitimating capacity will necessarily be limited. In terms of the typology of functions described at the beginning of this chapter, the inability to hear many cases reduces its effectiveness at all of them – the extent to which a Board might convince the public that the company is under control (propaganda), communicate to the public at large as well as to company decision makers problems with content moderation (informational), correct moderation mistakes, reform poor policies, or include diverse voices all scales with the number of the cases it can hear. That is, for each additional case it hears, that's a new chance to exercise public control over the company, learn and teach what's going on in the system, fix a mistake, exercise authority over a policy, or translate the voice of an underrepresented user or stakeholder (or Board member) into an outcome.

Yet as the number of cases the Board chooses to hear grows, its organizational challenges increase: A higher-volume Board may delegate more responsibility to staff, who may exercise undue influence over decisions, undermining its independence. Alternatively, it may need to designate subpanels to render decisions, which may undermine its consistency. The challenge of managing volume has, in other adjudication bodies, led to compromises in the authority of adjudicators for this reason.[6] Douek (2019, 6–7) has suggested – and she's obviously correct – that the problem of volume renders the Oversight Board incapable of providing something like individualized "due process" to users – instead, its function is to serve as a check on the general shape of the rules and their enforcement; yet at the same time, it's unlikely to be wholly effective in shaping company norms partly because of the difficulties of transmitting its results to "the globally distributed and time-starved workforce of content moderators that make the first instance content moderation decisions" and partly just because it lacks the "legitimacy" and "authority" to do so.

That being said, these workload pressures may still permit the Board to exercise the core function of insulating Meta from both internal and external pressure to deviate from its rules in the sorts of particularly high-stakes decisions where such pressures may be most threatening, at least to some degree. There are several preconditions for it to serve this function.

First, it must be genuinely costly for Meta to disobey its rulings; in particular, it must be more costly for Meta to disobey the Board's rulings than for it to obey them, even given the capacity of external actors such as disgruntled politicians to impose sanctions.[7]

[6] See discussion in Gowder and Plumb, "Oversight of Deliberative Decision-making: An Analysis of Public and Private Oversight Models Worldwide," Appendix E to Oversight Board global consultation report, https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf, 162–168; see also Klonick (2020, 2490).

[7] Douek (2019, 47–48) articulates a related, but, I think, mistaken critique. Drawing from an argument of Mark Tushnet's about the ineffectiveness of external checks on the powers of authoritarians, she

It is too early to assess whether such costs are available, but it seems to me that there is reason for concern in this respect, as such costs must be imposed either legally (i.e., through real government judicial sanctions on the basis of Meta's having violated its contract with the Board), or politically/economically (i.e., through public disapproval of company disobedience, and hence public response either through the political process, demanding more direct government regulation, or through the market, by abandoning the platform), and it is unclear that either avenue is readily available in the case of the Board.[8]

Second, workload considerations arise again to raise the concern that the Board may not be able to review enough cases to effectively serve as a constraining check in this sense under some circumstances. For example, if it can only review high-stakes individual cases, but external sources of pressure also care about low-stakes cases in the aggregate (e.g., pressuring the company to under-enforce hate speech rules against large numbers of individuals whose behavior has little individual impact but lots of impact taken together), then Meta may still be vulnerable to pressure in the kinds of cases that it cannot practicably delegate to the Board.

If the Board prioritizes selecting cases in which faithful compliance with company rules is likely to subject the company to costs that cannot so easily be inflicted on the Board itself – for example, cases involving powerful politicians and media figures – then it may be able to enhance its ability to serve a constraining function. Unfortunately, this doesn't address the problem of low-stakes cases that are high-stakes in the aggregate.

As to high-stakes individual cases, the Board ought to prioritize cases in which Meta might have an underlying temptation to break its own rules. That category includes those implicating the interests of external sources of illegitimate pressure, like demagogic politicians. It might also include the review of decisions to leave up content that is likely to be particularly profitable, such as that associated with high-revenue advertisers or popular content producers. For similar reasons, Meta ought to listen to Douek's (2019, 40–41) suggestion to provide the Board with review

argues that the Board "does not actually constitute a 'check' on Facebook's power [when] its actions remain in Facebook's best [long-term] interests." But this is an overly narrow view of what a "check" might be. It may be, and in Chapter 4, I argued that it is the case that the long-term interests of a company – like the long-term interests of Mancur Olson's stationary bandit – are aligned with those of the general public, while the short-term interests of a company are not. Under such circumstances it counts as a perfectly good "check" for the company or the dictator to have institutions that protect it from weakness of will, internal agency problems, and other kinds of pressures leading it toward short term decisions that conflict with the public good, in favor of long-term decisions that support it.

[8] In the case of political/market responses, company compliance would have to be sufficiently visible – either because of mandated company disclosure or some kind of post hoc investigatory powers by the Board. Moreover, the general public (or, perhaps, advertisers, who might have impact individually at a certain size) would have to care enough about compliance to coordinate on sanctions. I am uncertain whether the latter is the case. Moreover, ongoing controversies about things like Facebook's cross-check system, in which a journalist has alleged that Meta lied to the Oversight Board (Horwitz 2021), suggest that the capacity for monitoring compliance is likely insufficient.

authority over algorithmic recommendations and advertising – those are areas in which short-term financial temptations may lead the company to deviate from its rules.[9] Likewise, if programs like cross-check are allowed to continue, it may also be worth considering submitting membership in that program to Board review, as cross-check is essentially a list of people who get special solicitude because they're likely to be able to impose costs on the company.

Democratization, by conferring on ordinary people the capacity to participate in case selection and adjudication, could potentially mitigate the Board's workload-related problems by increasing effective staffing. For example, a multi-tiered system similar to American courts of appeal could be developed with regional popular adjudicators (such as a pool of users chosen from each country) serving as a first layer of appeal from day-to-day content moderators, whose decisions would be subject to appeal to the Oversight Board. Effectively, such a supplemental system could both introduce local knowledge to the adjudication process and serve as a form of workload management by refining issues and filtering meritless cases before reaching the Oversight Board. At the same time, it could create genuine deliberative opportunities closer to the front line of content moderation decisions, and thus both potentially improve uptake of the Board's decisions (since it would be easier to transmit them to intermediate appellate boards than to time-pressured mass workers), and deliver more protective process to individual users.[10]

Currently, the participatory character of the Board is present, but thin. It holds a fourteen-day public comment period for each case that it takes, giving the general public an opportunity to weigh in at will. However, this public comment process is likely to be subject to the standard weaknesses of public comment processes in other high-stakes environments, such as administrative agencies: Those with narrow interests are likely to have a stronger incentive to participate than members of the general public, the Board can largely choose what it does with the comments, and knowledge of the commenting process itself along with the skills to participate effectively are likely to be relatively elite resources. (This is especially so given the short two-week comment period – by the time someone who isn't plugged into

[9] Currently, the Bylaws (Art. 3, Sec. 1.1.2) contemplate future extension of the Board's authority to advertisements. Douek (2019, 42–44) also aptly raises a concern about nonremoval sanctions. It is unclear whether the Board has the authority, or will ever have the practical capacity, to review a variety of other kinds of "soft" sanctions such as reductions in visibility or demonetization. The risk of failing to carry out these expansions of the Board's authority is not merely that various forms of injustice to users or the public might go unreviewed, but that decision makers within the company might have an incentive to use one form of sanction as a substitute for another – to, for example, choose to reduce the distribution of some category of content rather than to take it off the platform – in order to evade Board review, and this might undermine the company's credible commitment to the Board as independent adjudicator.

[10] Of course, it might still be objected that the sheer volume of content moderation would overwhelm those intermediate entities as well. Even if that is true, they could nonetheless deliver *more* individual due process and introduce *more* contextual knowledge. One need not make the perfect the enemy of the good.

the Oversight Board process finds out about a case, it may be too late for them to comment.) Still, the comment process is better than nothing, and given that the Board makes comments public, it can have a potentially beneficial effect insofar as it facilitates scrutiny of the Board's reasoning process – ruling out, for example, clearly inadequate responses to public comments which could impair the Board's reputation.

In terms of the capacity to carry out reformist functions – or, on a more cynical story, to substitute its policy judgments for those of Meta personnel – in addition to institutional capacity issues noted above, there is also some degree of ambiguity as to the capacity of the Oversight Board to generate new rules of platform "law." Its charter provides that while the outcomes of individual cases are binding, the Board's policy guidance to the company is advisory, although formal policy advice will be addressed by the company.[11] Moreover, the charter provides that prior decisions "will have precedential value and should be viewed as highly persuasive" in "substantially similar" cases.[12] It's not terribly clear to me whether this is meant to be read as binding authority, persuasive authority, or something in between – but the Board's own practice may fill that out.[13]

Practically speaking, there are several ways in which Board decisions may have an impact beyond individual cases. First, the mechanisms for company response to policy advice may constrain the company by forcing it to give a reasoned explanation for its policies that can survive public scrutiny (cf. Klonick 2020, 2464). Second, to the extent the Board acquires in the future the institutional capacity to decide a large number of cases, it may constrain the company in practice simply by making rulings based on its own precedent. Third, it may influence company enforcement decisions to the extent those decisions are made "in the shadow" of subsequent rulings by the Board, especially if Meta is likely to suffer a cost from being scolded by the Board in some subsequent case for repeating the mistakes that the Board had already identified.[14]

The extent to which these sources of influence are effective will likely depend on several factors. How much external attention (and hence pressure) can Board decisions generate? How many cases can the Board effectively handle? In other words, if the Board decides to declare a new rule or interpretation, it must either be able to implement that rule/interpretation itself in future cases, or it must have sufficient

[11] Oversight Board Charter, Art. 4, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf; see also Oversight Board Bylaws Art. 3, Sec. 2.3 www.oversightboard.com/attachment/326581696050456/.

[12] Oversight Board Charter Art. 2 Sec. 2.

[13] Klonick (2020, 2463–64) reports that there was some disagreement in the design process on this question, which potentially explains the resulting ambiguity.

[14] To some extent, a formal capacity to generate precedent would also permit the Board to implicitly expand its institutional capacity, in the sense that decisions which it renders proposing major changes to Meta rules would have a broader effect on other cases.

sociological influence that the threat of scolding from it is meaningful to company decision makers.[15]

As a whole, we must evaluate the Meta Oversight Board as a promising, but limited, source of external constraint on Meta's decisions. The Board will never be as effective as an institution that genuinely empowers ordinary users to intervene on company decisions. However, it is likely to be reasonably effective in cases where stakes are extremely high, in which its decisions are most likely to draw the attention of regulators and the public at large, and in which there is the largest need for an external decision maker to check the unbounded authority of company personnel. In accordance with these suggestions, I will now turn to a direct examination of the Board's highest stakes case thus far – its decision regarding Donald Trump's indefinite suspension from the platform. Below, I argue that this decision demonstrated the Oversight Board's genuine potential.

5.2 A DEFENSE OF THE OVERSIGHT BOARD'S TREATMENT OF THE TRUMP CASE

Midway through the writing of this book, the event that seems to be becoming known as the "great deplatforming" happened – when Donald Trump, toward the end of his presidential term, was evicted from every major social media platform; at the same time, the notorious hard-right "free speech" social media platform Parler was also chased out of the Apple and Google app stores and even its Amazon hosting service. Everyone in the world knows why: The platforms had been used to plot an armed mob attack on the United States Congress aiming to stop the certification of Trump's election loss; Trump himself had made social media posts and speeches inciting that attack.

Trump's removal was also the Oversight Board's first major test, for Facebook's action against his account was submitted for its review. Fortunately, the Board rose to meet the challenge, affirming his removal in a public decision after receiving almost ten thousand comments from the public – but at the same time insisting that the removal not be "indefinite," and demanding a formal reconsideration of Trump's

[15] Thus, we cannot simply suggest, with Schulz (2022, 244), that Board interpretations of Meta rules amount to rule amendments. However, as I have suggested in the past (in a report for Facebook, no less), even adjudicative bodies without formal precedent-setting power tend to develop informal bodies of precedent. See Paul Gowder and Radha Iyenga Plumb, "Oversight of Deliberative Decision-Making: An Analysis of Public and Private Oversight Models Worldwide," report prepared for Facebook in the context of Oversight Board development process, distributed as Appendix E to Facebook, "Global Feedback and Input on the Facebook Oversight Board for Content Decisions," June 27, 2019, https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf, 172–3. It is notable that the Wikipedia ArbCom – the closest prior example to a platform court – seems, according to qualitative research that included conversations with some of its members, to have developed something like an informal system of precedent leading to at least some control over Wikipedia policies (Forte, Larco, and Bruckman 2009, 66).

removal after a time certain in order to reassess the danger posed by his continued access to the platform. To my mind, this illustrates a key function of post hoc rational reexamination of an emergency decision like the one undertaken after the horrifying events of January 6, 2021: The Board took an emergency exercise of (corporate) executive authority and disciplined it – preserving the protective act but subjecting it to an ongoing framework of rational determination under rules going forward. To see the significance of this, it will be helpful to take a small detour into theory.

For the infamous Nazi (yet still influential) legal theorist Carl Schmitt, the sovereign power of exception or of "commissary dictatorship" is a suspension of normal legal institutions necessary to preserve those institutions in the face of an existential threat (Schmitt 2005, 12–13; 2014, 118–19). This problem of emergency power frames Schmittian accounts of the sovereignty of political states. Because states are under an ever-present threat of emergencies that cannot be encompassed within their existing legal structure, sovereignty entails a kind of reserve capacity or prerogative to deviate from the pre-existing legal rules – to "decide on the exception." Typically, this entails the use of coercive force in some way or another – canonical examples include Lincoln suspending the writ of habeas corpus during the Civil War; Charles De Gaulle using the emergency powers granted by Article 16 of the French Constitution in the Algerian War; or the authority granted under Article 4 of the International Covenant on Civil and Political Rights to derogate from the other rights guaranteed by that instrument in states of emergency.

In the platform context, the Schmittian approach bears a striking resemblance to the "shock and exception" dynamic that Mike Ananny and Tarleton Gillespie have identified in an oft-cited conference paper, in which some terrible thing happens or some terrible platform practice is revealed to and criticized by the public (shock), leading to ad hoc exceptions made to solve the immediate problem (or take the heat off the company), but with no stable governance changes (Ananny and Gillespie 2017). We might consider the "great deplatforming" to be just such an example – at least at first.

While Zuckerberg, Dorsey, and their ilk manifestly did the right thing in chasing the insurrectionists and their leader off their communications tools – nobody has a right to speak directly to an armed mob attempting to overthrow a liberal democracy in their name, not unless they can absolutely guarantee that the only thing they'll say is "go home" – the great deplatforming also raised the tension about the idea of content moderation by revealing a quasi-Schmittian character at the heart of the enterprise of platform governance. Even if the leaders of the platforms ultimately agreed with claims, described in Chapter 4, that Trump's conduct had been violating essentially every platform's rules for a long time, until that moment the companies had not seriously acted to clean his pollution off the platform. Choosing that particular moment to chase him off, if understood as an application of those selfsame rules to be distinguished from the previous applications of those rules to keep him on, required an act of interpretation according to which on-platform conduct that

had not fundamentally changed between January 5 and January 7 suddenly assumed a different meaning in light of both Trump's off-platform conduct (in particular, his calls for "strength" and his lawyer's call for "trial by combat" at that rally) and the conduct of his supporters at the Capitol. Even the most conservative reading of that act of interpretation renders it *sui generis*: Only in view of the unique significance of Trump's speech could it be subjected to such an interpretative effort.[16]

Perhaps the key example of the major platforms taking no action (in Facebook's case) or taking much less serious action (in Twitter's case) before January 6 comes from an infamous 2020 tweet and Facebook post in which Trump threatened people who were protesting against police violence with military force. The context was highly conflictual summer 2020 protests over the police killing of George Floyd, protests that ultimately led to further violence, most infamously by one Kyle Rittenhouse (Sullivan 2021). In other words, the country was a powder keg, and the then-President of the United States took to social media to pour the following gasoline onto it:

> These THUGS are dishonoring the memory of George Floyd, and I won't let that happen. Just spoke to Governor Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you![17]

Twitter left Trump's looting/shooting post up but placed it behind an interstitial (Seitz 2021). Facebook, well, "raised concerns" and then Zuckerberg himself begged Trump over the telephone to tone it down. Ultimately, Facebook decided to bow to power and keep the post up – a decision reportedly made by Zuckerberg himself (Dwoskin, Timberg, and Romm 2020; Isaac, Kang, and Frenkel 2020; Swan 2020). In other words, Facebook's decision was an exercise of top-level executive power to keep up a post that, on any reasonable interpretation, was a direct threat to shoot protesters.[18] The contrast between the looting/shooting incident and January 6, and

[16] Arguably, the application of the rules changed with the relevant context, that is, with the fact that there was an ongoing violent attack on the US Capitol. This interpretation is supported by the fact that the Oversight Board decision on the Trump suspension attributes that suspension to rules referring to "events" that are violent and to "genuine risk of physical harm or direct threats to public safety." Oversight Board decision in Trump matter, www.oversightboard.com/decision/FB-691QAMHJ. However, as the Board also notes, the suspension was maintained beyond the duration of the attack due to ongoing threats of violence as well as "his continued insistence that Mr. Biden's election was fraudulent" – that is, conditions and rule violations not significantly different in kind from Trump's behavior prior to January 6th, as described for example in the Kamala Harris letter cited in Chapter 4.

[17] Bowden (2020), quoting the Twitter post, but the Facebook post was reportedly the same.

[18] The looting/shooting post seems to me to be much worse than some of Trump's statements during the attack on the Capitol. The latter at least did include a call for the attackers to go home, however insincere and self-undermining – for example, "This was a fraudulent election, but we can't play into the hands of these people. We have to have peace. So go home. We love you. You're very special. You've seen what happens. You see the way others are treated that are so bad and so evil. I know how you feel. But go home and go home in peace," the tail end of one of the statements quoted in the Board decision.

the direct intervention of Mark Zuckerberg in both cases, supports my interpretation of the "great deplatforming" as a kind of suspension of the existing (de jure or de facto) rules to deal with an emergency – possibly a company-threatening emergency, certainly an emergency threatening the overall liberal-democratic order in the United States in which the companies are embedded.[19]

That dynamic motivates my appeal to Carl Schmitt as a way of understanding what the deplatforming revealed. It's difficult to understand anything that happened with Trump (ironically a deeply Schmittian US executive) without the context of an executive-driven decision-making process responding first to external threats from potential regulators and then to much greater external threats associated with January 6. Ultimately, a state of exception was needed, and was declared.

But, as I said, the great deplatforming was undoubtedly necessary. When armed terroristic mobs are storming the Capitol, their communications must be disrupted in order to undermine their capacity to plan future attacks and to undermine the capacity of their leader to command (provoke? inspire?) such attacks. So, in a book like this, which proposes the importation to platforms of organizational strategies from democratic and lawful governments – strategies that are fundamentally anti-Schmittian – the observed need for an act of sovereign heroism stands as a fundamental challenge. Dare we subject the platforms to internal law and to popular control? Could platforms subjected to internal law and popular control have kept Trump away?

5.2.1 *How the Oversight Board's Trump Decision Serves as a Counterexample to Carl Schmitt*

Unexplored by Schmitt's theory is what happens after the executive act declaring the state of exception and responding to the emergency is complete. I contend that there are circumstances according to which emergency executive action might provide feedback to the overall system of rules and support, rather than undermine, something like the rule of law – for both polities and platforms. Beginning with the state context – I would suggest that considering the aftermath of an exercise of emergency power somewhat turns Schmitt on his head. For raw emergency executive power is self-undermining just because, as described in Chapter 4, the rule of law is itself necessary for effective exercise of power. Executives making use of emergency

[19] By the notion of *de facto* rules suspended in the great deplatforming, I mean to suggest that the fact that the major platforms were ignoring – or at least over-charitably interpreting – their own rules by failing to do anything about Donald Trump for years beforehand amounted to a kind of effective law-on-the-ground giving high political leaders a different set of rules. Interpreted generously, this parallel set of rules was made on the grounds of some analogue to "newsworthiness," or the importance of public visibility into the words and actions of their leaders; interpreted cynically it was motivated by the profitable engagement that Trump's behavior generated and the fear that he and his allies would engage in regulatory retaliation otherwise. As noted in Chapter 4, similar informal policies had evidently been applied to powerful politicians in India and Brazil (Purnell and Horwitz 2020; Marantz 2020).

power need some reason to believe that their commands will be carried out, and that belief in turn depends on a broader institutional context in which they can do things like make costly threats credible. *There is no power without the capacity for constraint.* Hence, emergency powers, to be meaningful, require some way of, in effect, regularizing their use, at least after the fact.

I contend that courts and even quasi-courts like the Oversight Board are suited to carrying out such post hoc regularization. With respect to courts – in countries with functional judiciaries, emergency uses of executive powers tend to be subject to challenge after (and sometimes during) the fact. But the core modality of judicial institutions is reason-giving: The thing that makes a court a court, as opposed to some other kind of authority, is that it states general rules and explains why a given use of power over an individual is justified by those rules. Common-law-style courts, that is, courts that generate precedent which itself counts as an authority in future cases, have the further capacity to apply those rules moving forward.

The confrontation between a court and an executive having exercised emergency power is therefore generative. Whether that court upholds or overturns the use of emergency power, a court can attempt to articulate the bounds of that power, and its criteria for application, with reference to the specific facts of the emergency that was presented by the dispute. In doing so, at least sometimes the court can articulate the rules under which similar acts might be permissible in the future. Such an action can, in effect, bring future emergencies of the same form within the system: The next time a closely related threat appears, the executive may not need to suspend the rules to address it, but may be able to follow the rule laid down by the court in the wake of the last emergency. In effect, a court can rationalize and normalize emergency action on a forward-looking basis.

Sometimes, this judicial power can even be self-limiting in a dialogic fashion. For example: One of the most infamous and rightly condemned decisions of the United States Supreme Court is the Korematsu case, in which the Supreme Court upheld the race-based internment of Japanese-Americans during World War II.[20] The case is deeply evil, and it was finally reversed (in dicta) in another evil, much more recent, case, *Trump v. Hawaii*, which upheld Donald Trump's notorious Muslim Ban.[21] Yet Korematsu in part also represents a kind of domestication of emergency power, for while the case permitted the President to carry out the internment, it also was one of the earliest and most important of the articulations of the "strict scrutiny" standard for judging government race discrimination; a standard that later civil rights organizations could use to argue that race-based government action was "presumptively void" (Robinson and Robinson 2005). This is not, of course, a defense of Korematsu – the case was an abomination against justice. But rather, it's a defense of what a scholar like E.P. Thompson (1975, 258–69) or Lon Fuller (1978, 365–81)

[20] *Korematsu v. United States*, 323 U.S. 214 (1944).
[21] *Trump v. Hawaii*, 585 U.S. ___ (2018).

would identify as a kind of valuable normative character of the process of lawlike adjudication: Even a wicked act, when it is filtered through judges or quasi-judges looking to state rules that respond to reasons and make an act compatible with a legal order, can carry within itself the seeds of its own reform.[22]

This is, I contend, part of what happened – at least *in potentia*, depending on the long run response to the Trump decision within Meta – with the Oversight Board's response to the great deplatforming. Zuckerberg's decision was submitted to the Board, and the Board helped regularize it by articulating the principles that justified the executive action and further integrated those principles into the presumptive guidelines of the platform going forward – so that the next time similar threats arise, they can be accommodated without declaring a state of exception. The Trump case thus illustrates how an entity like the Oversight Board can fit into, and help alleviate, the tension between the Schmittian character of company leadership (especially in cases of emergencies) and the value of self-constraint understood as an analogy to the rule of law.

Observe that the Board's decision on the Trump suspension seems to recognize, at least in part, the emergency character of Facebook's act. While there's a certain lack of clarity to the decision, and specifically to the extent to which Facebook appealed to the uniqueness of the situation in justifying the suspension (or was merely motivated by that uniqueness), there are references not merely to Facebook's pre-existing policies on "Dangerous Individuals and Organizations" and incitement, but also to the need to preserve a peaceful transfer of power in the United States, and hence implicitly to the context of the threat to that peaceful transfer occasioned by an attack aimed at preventing the certification of Trump's electoral loss. Moreover, it is quite clear that the Board perceives the particular sanction imposed – an indefinite suspension, as opposed to a time-limited suspension or full-fledged account deletion – as *sui generis*. With respect to that sanction, the Board expresses some sympathy for the exigent circumstances involved, and, while disapproving of the uncertainty created by the indefinite suspension (which grants excessive discretionary power on an ongoing basis), approves of a very similar process: Time-limited suspensions that are renewable, upon the reasoned conclusion that the ongoing risk of incitement continues at the end of the initial suspension. Effectively, this is

[22] In a weaker sense, a bureaucracy tends to generate internal rules, policies, and procedures to implement an executive command. To bureaucratize such a command is to set up structures of authority and rules to generalize it and apply it across the administered domain. For Weber, bureaucracies are forms of "juristic" or legal authority, which share with the law the appeal to general rules to justify their actions (Weber 1946, 299). Translated into the executive power, this distinguishes two kinds of top-level executive commands: The command "go do X to Y" ("go shoot that dissident," "go ban that particular troll from Twitter"), which is a one-off act, and the executive command "go establish and implement a policy of doing X to Ys" ("shoot all dissidents," "ban all trolls"). The latter, even if issued arbitrarily to respond to an emergency, may have a rationalizing function, as at least *future* cases of similar threats will be subject to being addressed under existing rules as opposed to *sui generis* acts of executive power.

a procedural gloss on the indefinite suspension that requires the company to revisit its decision on a periodic basis (but where the period itself appears to be in the company's discretion, or at least not discussed by the Board). This revisiting, however, must be *reasoned* and hence implicitly subject to review and disagreement like any other form of rule-bound action.

This corresponds fairly well to the post-Schmittian framework I have outlined: Even though the Board acknowledges the emergency nature of the suspension and that at least some of the company's rules were derogated from in the process, it both retroactively justifies the basis for the suspension in terms of pre-existing policies (while providing recommendations for the clarification of other policies that may have been applied arbitrarily), and reforms the actual sanction imposed on an ongoing basis to be more compliant with law-like norms without undoing the resolution of the emergency.

Thus, the capacity for Facebook's executives to respond to the emergency was preserved, as was the actual effect of the action: Facebook wasn't ordered to give Trump his account back. In that sense, it held onto the benefits of the Schmittian executive. At the same time, the Board sketched an outline for future responses, not just to Trump but perhaps to individuals in similar positions more generally (consider that the world presently faces parallels to Trump in other nations with massive Facebook user bases, such as Jair Bolsonaro and Narendra Modi). It did so, in effect, by articulating the implications of Facebook's existing commitments, both in its own rules and in its statements about human rights, to such cases.

Moreover, it fills out those commitments in a way that supports the ambiguous claim to precedential power in its founding documents. For a key example, the Board notes that Facebook applied the principles from the Rabat Plan of Action, a standard for considering incitement to hate developed by the United Nations High Commissioner for Human Rights. Although the Board does not, formally speaking, generate precedent that binds the company – just itself, and weakly, as noted above – the fact that the company already applied the criteria from this internationally recognized human rights framework, and the Board explicitly approved of it, suggests that it could be the basis for a kind of informal system of Facebook caselaw, insofar as decision makers within the company sorting out what to do with the next instance of serious incitement are likely to recognize that relying on the Rabat Plan is more likely to ensure that their actions will be upheld.

Critically, the Board's condemnation of "indefinite" suspensions may actually facilitate, rather than restrain, Facebook's capacity to control the behavior of powerful political leaders on its platform. The problem with an indefinite penalty is that it can be revisited at any time – and thus it puts the people with the authority to revisit that penalty in a strikingly weak position with respect to resisting the pressure of powerful political groups. Until the Oversight Board decision, there was no internal basis for the company to say, to an angry Trump-linked pressure group, "no, our policies require that we only reconsider the case at the following date certain [X],

and at that time, you will be required to demonstrate the following things [Y] in order to show that Trump can return to the platform consistent with the safety and human rights interests underlying our rules." The framework offered by the Board – if implemented by Meta and backstopped by some real sanctions for company noncompliance – would provide just such a basis. In response to pressure groups, Zuckerberg or other executives could offer a neutral reason – compliance with the Oversight Board's command to regularize the terms of Trump's suspension – for considering letting him back on the platform only at a certain time and for bounded and relevant reasons. The fact that such decisions when made will be reasoned and subject to further Board review can further support the message that the decision to maintain Trump's suspension was an act of rule-following, not partisan bias.[23]

5.2.2 *Can Platforms Have a Constitutional Identity?*

There's a sense in which this idea of rationalization is latent in Schmitt's conception of sovereignty, at least as transposed to liberal states and, bluntly, denazified. For the root of sovereign power on Schmitt's account is an identified and bounded people on behalf of whom the sovereign acts. This is what leads to Schmitt's (1996) notion of the friend-enemy distinction: A state as a bounded group depends on the notion of "the political," which in turn is understood as the capacity to point outside and say "these are our enemies" by way of contrast. But pluralistic liberal states tend to be ambivalent at the least toward the notion of defining a people with reference to its enemies, especially after the German home of Schmitt's theories showed where they could all too easily lead. To avoid the dangers of such nationalism, contemporary theorists associated with the idea of "constitutional patriotism" such as Habermas (2001) and Jan-Werner Muller (2007a) have suggested that the people can, in essence, be defined in terms of its legal system and the commitments that system represents toward an ongoing enterprise of legal self-definition.[24]

I contend that we can make sense of platforms as having something like a liberal-democratic legal identity in two senses. First, the interests as well as the more aspirational and clearly articulated organizational missions of platform companies require a framework of functional liberal-democratic political states. That gives the companies a reason to defend the boundaries of that politics and to deny the use of

[23] While this book was in production, Meta announced that Trump's account would be restored. Nick Clegg, "Ending Suspension of Trump's Accounts With New Guardrails to Deter Repeat Offenses," Meta Newsroom, January 25, 2023, https://about.fb.com/news/2023/01/trump-facebook-instagram-account-suspension/. The company's announcement is consistent with the framework described in this chapter, in that Meta acknowledged that Trump's suspension was an emergency act but described reforms to its rules and to the sanction imposed on Trump undertaken to regularize the situation in accordance with the Oversight Board's ruling. The announcement also set out (albeit briefly) reasons for restoring Trump to the platform (as a product of the company's evaluation of the ongoing risk) and specific policies for addressing Trump's behavior going forward.

[24] See further discussion in Gowder (2019, 349–54).

their platforms to those who would destroy the normal liberal-democratic constitutional order. Second, the practical imperatives of successfully operating platforms that are governable in cross-national contexts are, for the reasons described in the rest of this book, dependent on at least something resembling liberal-democratic institutions in a minimal sense internally, that is, processes for the participatory exercise of reason.

Both the ideologies of the employees and the founders of those companies and the practical conditions for their functioning as businesses assume that they're mostly operating in liberal democracies, or at least that their employees can work out of liberal democracies when the angry government officials come to try to force them to regulate content in ways contrary to their own interests. This can be seen, for example, in the First Amendment defenses that the companies rested on in response to legislative efforts to regulate their content moderation practices in Florida and Texas: In order to keep from being forced to host extreme right hate speech that would drive off their users, they need to be able to defend their own editorial independence.

More abstractly, the notion of a many-to-many networked platform only makes sense in the context of an overall view of the social world which in the first instance conceptualizes individuals, qua users, *as* individuals, that is, as capable of deciding for themselves which associations to engage in and of building multiple layers of association representing as well as crossing between different relationships in which they stand with one another (as consumers, as producers, as co-citizens, etc.) – which, in other words, sees the individual as prior to their existing affiliations and sees those affiliations as contingent and mutable. Understood as liberal in this sense, it should be no surprise that the stated ideology of every major platform has trumpeted its commitment to individual freedom and choice (e.g., Adams and Kreiss 2021, 40–57; Halliday 2012).

In the context of the rationalizing function of law-like adjudication, we might also suggest that a platform's identity is partly constituted by its rules and what it does with them. To be sure, as I have emphasized at multiple places in this book, the rules cannot be cleanly distinguished from the overall affordances a platform offers. The definition of the kinds of activity that can be carried out on the platform, either in a positive sense ("here are what the tools on offer are") or a negative sense ("here is what you cannot do") is a fundamental part of the value proposition of such a platform – and thus, in a capitalist environment, also a company's identity. For a more concrete example, part of the way that Facebook and Twitter are different is that Facebook has its famous "real name" policy, which shapes the kinds of interaction that people expect and experience on the platform, and Twitter does not. It's a different kind of (virtual) space, with a different kind of *telos*, setting different kinds of expectations in pursuit of different kinds of goals.

Connecting those ideas to adjudications like the Trump case in the Oversight Board: When rules are articulated in response to emergency executive acts, we

can interpret the series of events as a kind of practical self-learning by a platform, which fills out a vision of its product-userbase-personnel-society nexus, and hence its identity.[25] In turn, the articulation of that identity potentially makes up a normative defense of the initial executive act. In concrete terms: The legalistic or bureaucratic rationalization and domestication of the great deplatforming could provide a retroactive justification of that act, as an interpretation of it in terms of a commitment to preserve the liberal-democratic order in which the platforms are embedded and invested.[26]

Perhaps, however, I ought to offer a more involved defense of the notion of platform identity, for I believe it might serve a broader function that can contribute to the resolution of some of the other problems articulated in this book.

5.3 CAN PLATFORMS HAVE A LIBERAL-DEMOCRATIC IDENTITY?

The best way to begin thinking about platform identity is to start with the notion of "free speech."

One of the classical challenges in liberal-democratic politics is the problem of "tolerating the intolerant." Simplifying a little bit: Liberal societies have a commitment to values like free speech and the marketplace of ideas. But does that commitment extend to those who espouse the denial of those same ideas? Can a society permit its free speech to be used to promote a censorship campaign; more starkly, can a liberal-democratic society with free and open elections permit candidates of a political party with the espoused ideal of overthrowing the democracy and implementing a dictatorship to stand for office?

Political philosophers have struggled with these problems for generations without making a lot of progress. In the meantime, however, the terrain of the problem has expanded from governments to companies, and especially to social media companies. Those companies typically are founded and run with a commitment to ideals of "voice and free expression" (Facebook 2019), or have declared themselves as being "the free speech wing of the free speech party" (Halliday 2012), or have stated

[25] For the philosophers in the audience: This is self-consciously Hegelian, I admit it – but you should have seen that coming back in the introduction when I appealed to Dewey's very Hegelian democratic learning framework. I make no apologies.

[26] On the possibility of retroactive justification, see Gowder (2019). Such a view, incidentally, turns Schmitt on his head, or, perhaps, twists him beyond recognition: In the presence of a functioning system of post hoc rationalization, the distinction between commissary and sovereign dictatorship first disappears, as the acts of the commissary dictator themselves become partly constitutive of the postcrisis normative order; then the state of exception is permitted to gradually disappear, or at least shrink as a liberal legal order encompasses an ever broader set of possible crises; finally "the political" itself dissolves under the pressure of a conception of identity that focuses not on groups of people but on sets of acts. While this book cannot explore the prospect of reading the experience of conducting platform governance back into political theory in order to unsettle or outright overturn conceptions of sovereignty and state like Schmitt's, I observe here that the capacity to do so is yet another reason to bring together the study of politics with the study of platforms.

as a core value "that everyone deserves to have a voice."[27] And this is not merely a value commitment but also an economic one: Because platform economic models are inseparable from chasing positive network externalities, the default strategy is to operate with an expansive conception of the addressable market (userbase); moreover, hosting a truly wide variety of people (with their associated beliefs, interests, and goals) creates a key advantage of large-scale social media platforms in particular, namely, the immense diversity and hence the immense capacity that such platforms have to facilitate niches for people who otherwise would be unable to benefit from the sociality gains to the network.

Even in liberal states which are in principle committed to a kind of neutrality among conceptions of the good (Patten 2012), their governments still frequently themselves adopt controversial positions as a product of public value. A famous statement of this idea in United States law comes from the Supreme Court case of *Rust v. Sullivan*, which upheld, against First Amendment challenge, a government funding scheme for family planning clinics that restricted recipients from offering abortion services using public money. In the pithy words of Chief Justice Rehnquist: "When Congress established a National Endowment for Democracy to encourage other countries to adopt democratic principles, it was not constitutionally required to fund a program to encourage competing lines of political philosophy such as Communism and Fascism."[28] While a debate about the appropriateness of such a statement is possible, the notable point for present purposes is its resemblance to the notion that the United States has a public identity that is inconsistent with support for communism or fascism.[29]

Platform companies tend to be inclined to offer their resources even to advocacy for views that are radically inconsistent with company values: While there are some limitations associated with things like hate speech rules, and there is some reason to think that platform content moderation is moving broadly in the direction of

[27] YouTube, "About YouTube," https://about.youtube/ ("Our mission is to give everyone a voice and show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share and build community through our stories").

[28] *Rust v. Sullivan*, 100 U.S. 173, 194 (1991) (internal citation omitted). Of course, to the extent Rehnquist meant to equate abortion with something like fascism, I do not endorse that implication.

[29] In terms of constitutional law, and the theoretical implications of US law for the idea of liberal neutrality, I am radically simplifying matters for the sake of illustration. In reality, the United States is constitutionally obliged to offer some of its resources even to advocacy for communism and fascism, in the form of public forums – areas of government property, such as parks and streetcorners, traditionally held open to all comers. Also, another way – although an oversimplified and problematic way – to think of the problem in liberal states is that liberal neutrality is in terms of *conceptions of the good*, that is, individual values and goals, not neutrality with respect to the political ideologies enabling or threatening the stability of the institutions that make liberalism possible. But, of course, some conceptions of the good entail stability-threatening political commitments (e.g., various theocratic religious views like integralism), and *democratic* states may also be obliged for democratic purposes to tolerate a wider range of attacks on their fundamental modes of political organization. At any rate, the complexities here run extremely deep, but the comparison to states is merely for the purposes of providing a cognitive entry point into the options for platforms, and I don't aim to contribute to the literature on liberal or democratic neutrality/toleration here.

acknowledging broader political commitments and dependencies (e.g., Horwitz and Scheck 2021), as a whole, the baseline view of most platforms seems to be that even, for example, advocacy for political ideologies that might entail the destruction of the platforms themselves is fair game on those same platforms.

Yet the problem of tolerating the intolerant as it's manifested in the political context generates a kind of internal tension that also applies to platforms. Free speech and toleration as core commitments can carry the seeds of their own destruction. On social media, this problem burst into the public eye for the first time in the "gamergate" crisis back in 2014, which provided the world with the spectacle of many prominent women being driven offline by campaigns of extreme gender-based harassment. In other words, the capacity of social media spaces to host feminist gaming voices, and hence the capacity of feminist gamers to engage in free speech, was seriously undermined by the misuse of the free speech of others. For one particularly clear example, one gamergate technique was "doxing": revealing the personal information of their victims publicly, and thus making it possible for criminal third parties to threaten them with violence for their speech.

Part of this paradox is built into the nature of the "free speech" ideal itself. Speech is not simply additive. Some speech can destroy other speech. Gamergate is one example. America's long struggle with electoral campaign finance is another: While the US Supreme Court has repeatedly held that campaign spending is a form of free speech (buying advertisements costs money), critics of the money=speech equation have long pointed out that those rulings permit the wealthy to dominate politics, and thus effectively stifle the voices of smaller and poorer groups. For platforms as for states, so long as one person's speech can be used to capture, undermine, or destroy the conditions necessary for another person to speak freely – whether that's through whipping up the threat of violent retaliation from third parties or just buying up every possible way that anyone else might get access to listeners, readers, audiences – anyone who wishes to run a forum where free speech is genuinely possible for all must curate an ecosystem in which the affordances of speech are genuinely universally available, and speech-destroying-speech is itself excluded.

Because of a growing recognition of this property of free speech, eight years on, the right response to gamergate seems fairly clear to most of us: The companies should have acted much more vigorously, within the limits of their capacity, to protect the victims of this harassment. The leading scholar of online harassment, Danielle Citron (2019), has given a history of the growing recognition of the expressive harm of permitting online harassment. As she explains:

> Cyber harassment destroys victims' ability to interact in ways that are essential to self-governance. Online abuse prevents targeted individuals from realizing their full potential as digital citizens. Victims cannot participate in online networks if they are under assault. Rape threats, defamatory lies, the non-consensual disclosure of nude photos, and technological attacks destroy victims' ability to interact with others. They sever a victim's connections with people engaged in similar pursuits.

> Robust democratic discourse cannot be achieved if cyber harassers drive victims away from it. Victims are unable to engage in public dialogue if they are under assault.[30]

But we have not yet interrogated the underlying normative standpoint from which that new consensus is to be articulated – US free speech norms? Global human rights? Some combination of those things? After all, the conception of what sorts of speech causes harm to the speech interests of others is not measurable by some neutral criterion – we can't simply count the total number of words uttered and conclude that the speech regime in which that number is highest is, therefore, the best (if nothing else, vicious harassers can be *prolific*). It's doubtful that "better for speech" could be measured at all, but if it can, any measure adopted would necessarily make reference to an underlying value to be promoted, because the judgment in question is ineluctably normative.[31] In other words, platforms, like states, have to come to some kind of position on what "free speech" *is* in order to promote it. The very idea is underspecified, and can only be filled out with respect to a thicker set of goals – consistently pursued, an identity.

There is precedent for the notion of a thick organizational normative identity in this sense. Platforms bear at least a substantial (if not complete) resemblance to some of the larger transnational complex organizations that have vexed political theorists since the Middle Ages. Consider the medieval Roman Catholic Church, or the Knights Templar, organizations that transcended international boundaries, yet exercised their own kind of quasi-legal influence over their members and were (much like platforms) perceived as threatening by local rulers partly in virtue of those facts (Levy 2017). Those organizations are key examples of what Jacob Levy (2017), in his canonical study of the phenomenon, calls "intermediate associations," which have traditionally posed a challenge for theories of government centering on states.

The worries about the relationship between such intermediate organizations, states, and people, are remarkably similar: Levy aptly diagnoses a persistent tension between the risk that intermediate organizations will gain power over their members which in turn deprives them of the rights associated with liberal democracies (if you're a conservative, think of university hate speech regulations; if you're a liberal, think of churches with retrograde views on gender roles), and the countervailing risk

[30] Citron (2019, 130).

[31] For example, we might say that a world with more speakers, or more diverse speakers, or more representation of socially subordinated speakers, is better. But that implies an underlying value judgment about which kinds of speech, and from whom, we value. Missing this point is the core error of the US Supreme Court's campaign finance jurisprudence following *Buckley v. Valeo*, 424 U.S. 1 (1976): The notion that restricting campaign finance is a restriction on speech ignores the fact that, among other things, spending on campaign finance is embedded in a competitive market where one speaker can potentially outbid others, and hence that a lack of campaign finance regulation amounts to a *de facto* restriction on the speech of less wealthy interests – a controversial interpretation of the notion of "free speech" that accepts a tradeoff between an increase in quantity of speech for some in exchange for a reduction in the number of speakers.

that excessive state control over such organizations also deprives individuals of the capacity to exercise those freedoms in association (including by waiving them in the pursuit of shared ends). Compare a persistent worry about social media: Does government control over social media content moderation (like the Texas and Florida laws)[32] protect users' freedom (particularly political freedom) or does it undermine it by impairing their ability to choose to participate in communicative ecosystems under known (if restrictive) terms?[33]

Yet an important point that Levy emphasizes about such organizations, which distinguishes them from liberal states, is that they are purposive: They represent shared organizational ends and hence can be expected to generate (and enforce) behavior that varies from what would be predicted under pure liberal neutrality in pursuit of those ends. Universities are dedicated to the production and sharing of knowledge, churches are dedicated to the spiritual well-being of their members, and so forth.

Social media platforms, like other commercial entities, aren't quite as closely connected to an organizational purpose. There wouldn't be much of a point to joining a noncommercial association without some sense of organizational purpose – the reason one joins a church or a university is, arguably, just to participate in that shared end. By contrast, there may not be a shared end in the same sense to, say, Twitter – the company's end may just be to make money, and individuals may join it simply for individual transactional benefits.

Yet the same can also be true for more conventional kinds of intermediate associations – students may join a university just for a credential in the job market, faculty may join a university just for a cushy job with high job security and autonomy – and indeed, universities have been criticized for a long time for putting financial imperatives ahead of their intellectual mission (the reader who teaches at a large American university may simply reflect on the existence of your football team).[34] Moreover, the commercial character of a platform might also give rise to a second-order organizational purpose to the extent that its revenue model is attached to the instantiation of particular values.

[32] Florida: S.B. 7072; Texas: H.B. 20.

[33] The parallel with the Catholic Church in particular (as opposed to other kinds of medieval-era organizations such as the Knights Templar) is striking because a notable fact about the church is that it generated its own ecosystem of intermediate, and partly autonomous, organizations relative to itself – monastic orders, the Society of Jesus, and so forth, much like Facebook or (especially?) Reddit itself spawns distinctive groupings of people. The Catholic Church also developed its own canon law and courts, much like Facebook.

[34] This is also true of historical examples of intermediate associations. For example, Levy offers the Knights Templar as a core example of an intermediate association But a part of the reason for the destruction of the order was that they'd effectively turned into a banking organization rather than an organization devoted to Christian chivalry and whatnot – a character that became rather inconvenient for them when the King of France needed their money and their debtors saw no particular reason to come to their defense (Nicholson 2021, 73, 80). About the medieval Catholic church and the constant battles over slippages from its religious character from the Constitutions of Clarendon to the Reformation, of course, little needs to be said.

Indeed, that is ultimately the source of social media platforms' commitment to free speech. They have not just an ideological commitment to ideas of free speech arising from the libertarianism of their corporate founders, but an economic model that depends on the promotion of diverse communicative content. And promoting diverse content requires a difficult balance between permissiveness and control to permit diverse *people* on the platform. That is, the core interest of the major social media companies is not just "free speech" in the abstract. Rather, it is free speech in concrete terms of a healthy ecosystem in which people are actually capable of connecting with those who share their goals and interests – and that concrete implementation of the ideal of free speech requires defending the ecosystem against the kinds of toxicity that can drive out beneficial diversity. Social media platforms are markedly similar to universities in this sense: While universities ought to be sites for daring intellectual debate, their mission to promote goods like student learning also rules out giving free reign to kinds of communicative interaction that impair that learning mission.[35] Hate and censorship are *both* anathema to such an ecosystem.

In effect, social media companies have economic – as well as ideological – reasons to run their platforms to promote ends similar to those of a sophisticated – not a naive – liberal democracy, and to promote the flourishing of autonomous groups of their users. At the same time, their well-being is also tied in with the success of liberal democracy in the external world. Repressive states rightly view social media as simultaneously a threat (to the extent they provide their citizens with affordances to organize against them, *a la* the Arab Spring) and an opportunity for surveillance and manipulation (e.g., to the extent pernicious propaganda can reach its victims through the platforms, *a la* Myanmar, the Philippines, Trump, and many other examples). So at best such governments attempt to subvert platform content moderation efforts or coerce them to become tools of the regime, at worst to ban them entirely. And being a social media user in a repressive state, unless one happens to be organizing to overthrow it, is likely a bad idea in view of the likelihood that one's government will abuse platforms to surveil and manipulate one.

With respect to the largest social platforms (and here I mostly speak of Meta/Facebook and Alphabet/Google), in some countries, a platform's userbase effectively *is* the public (e.g., Facebook in Myanmar). To the extent that a platform constitutes a significant part of a country's public sphere, promoting sophisticated liberal-democratic interactions on the platform also amounts to promoting them in the public at large, and hence the goals of providing a welcoming and diverse

[35] In the university context, an instructive example is the case of University of Pennsylvania law professor Amy Wax, who misused her free speech rights to make public comments on the alleged intellectual deficiencies of her Black students (Chotiner 2019). Such speech goes well beyond the bounds of academic freedom because it unavoidably indicates to some of her students that they cannot expect equal treatment in a professor's pedagogical function.

communicative environment and of supporting liberal-democratic institutions can align.[36]

I thus propose that we call this combination of value and economic commitments – which may in part be determined by its rulemaking and adjudicative process – a component of a platform's "identity." We might also use the word "mission," although I prefer "identity" not just because it avoids the cognitive sludge associated with corporate propaganda documents like "mission statements" but also because "identity" as opposed to "mission" captures a sense of the present; not of a thing to be aspired to but a thing to be and to preserve. Unlike mission, "identity" also reflects the recognition of the past: The identity of an entity isn't just something decided on as a goal to achieve but also something constituted by an entire history of behavior.

In this sense, the notion of identity can operate bidirectionally in shaping and rationalizing (in the positive sense) a platform's day-to-day decisions when those decisions are controversial or uncertain.[37] Looking forward, the notion of identity can combine a sense of mission and an appeal to consistency, allowing, for example, a significant content moderation or design decision to be guided by reflection on the way that its outcome can be integrated into a platform/ecosystem/firm self-understanding drawing from past decisions, future goals, economic self-interest, and social values.

The attentive reader may have noticed an equivocation in the previous paragraph, helpfully denoted by a couple of slashes: Is a platform identity a property of a firm or of an entire ecosystem? The answer may be both or either: that is, we might imagine a firm having an identity derived from the views and behaviors of its management, employees, shareholders, and so forth; or we might imagine a broader quasi-*demos* composed of a firm in its social context and its users having such an identity. The best characterization of an identity will depend on the behavioral and organizational context: To the extent a broader group of people shape the key decisions and affordances of a platform – for example, if those decisions are rooted in participatory input from a broader userbase – it will make sense to understand a platform's identity as incorporating those users as well.[38] This is simply an implication of the notion that identity is backward-looking, determined not just by someone deciding "this is our identity" but also by reflecting on a pre-existing course of conduct.

Consider how this idea might interact with a concrete example like gamergate. The first beneficial effect of a focus on identity is a shift in the locus of concern from the competing interests of a bunch of external agents – the demands of various

[36] Of course, we must not forget that this could be understood as a pernicious form of colonialism, but for present purposes, I ask the reader to bracket this idea – I'll address it in a moment – and think simply about the notion of platforms coming with their own ideological commitments.

[37] Douek (2019, 49–50) articulates a similar idea, aptly observing that "Facebook cannot escape the need to make a choice about the kind of platform it wants to be."

[38] Since this book ultimately advocates expanding the number of people who get a say in platform company decisions, it also by extension advocates expanding the scope of platform identities. That isn't an accident.

individuals for a microphone or for protection against harassment – that is, "content moderation" – to what we might call "curation" instead, viz., the notion that there is a counterfactual ideal (or at least better) state of affairs on the platform, represented by, for example, a characteristic way that people interact and experience the communicative ecosystem. In terms of concrete responses to particular situations, it's easy to think that 'what do we want our ecosystem to look like' could have generated a swifter and more effective response to gamergate harassment simply by foregrounding the platform-wide interest in not driving away valuable contributors, particularly when those driven away are members of a group underrepresented in their socioeconomic context (female video game players and developers) whom social media could enable to flourish (and who in turn could enrich social media).

More broadly, the concept of a platform's identity reopens the problem of political bias which was a core obsession of Chapter 4 in the form of claims by both the left and right in the United States that platform rule enforcement is biased against them. There's an inherent non-neutrality to platform rules. Consider the prohibition on hate speech: that's a value judgment that certain social media companies have made (albeit at least in part also as a consequence of an economic judgment that their users and advertisers don't want to see that stuff). And it's a value judgment that is inconsistent with certain major political positions, including what might potentially be mainstream positions in certain societies, at least judging by electoral results. Donald Trump said any number of things in public during the 2016 elections that would qualify as hate speech under Facebook's rules, and the notion of platform identity gives companies some reason to wear that fact on their sleeve and shrug off claims of "political bias."

Platform identity is also useful in a second way which responds to the worries raised by Chapter 4: A governor with a clear identity is more predictable to the governed, and more capable of resolving internal conflict, as well as making ongoing decisions going forward about novel situations. In other words, it is more capable of supporting something like a rule of law.

5.3.1 *Toward Participatory Platform Identity*

The notion of platform identity may seem to re-awaken the worries about colonialism associated with efforts to export liberal ideals to other kinds of community. For example, would the notion of a liberal Facebook or a Twitter identity really be a just guide to governance decisions when a company operates in a Muslim country? And are the self-determination interests of the users in that country really to be subjected to a kind of ideology represented by a platform that is itself framed in self-consciously American terms?

There are at least two answers to this objection, one less convincing than the other. The less convincing idea appeals to a kind of transparency ideal represented by markets. To the extent that a company wears its ideology on its sleeve, the existing

government, as well as the public of a state, can identify that ideology and make a relatively free choice in a market context about whether to permit or participate, respectively, in that ideology. As stated, this response is wholly unconvincing, for, we know that market choices like that are not free in as robust a sense as conventional American ideology would have it (e.g., because of the endogeneity of market preferences, see Satz 2010, 180–81), and indeed the incentives offered by platform companies to people in the global South to use their platforms (such as Facebook's "Free Basics") could easily be seen as a *component of* rather than a *response to* colonialism.

A slightly more convincing version of the first response is that the notion of identity may actually serve as an input to some of the questions surrounding the international involvement of platforms, particularly in repressive countries. To the extent liberal democracy is integrated into the identity of a platform, it may be reasonable to just say that its presence in certain countries – particularly those countries in which a company may experience itself as forced to choose between deferring to the decisions of an illiberal government and substituting its own values for those of a local populace – is inappropriate. It might be that operating in such a country – at least without a way of integrating the people of that country into the company decision-making process directly – is incompatible with being a liberal-democratic platform, and hence the company should simply stay out.

But this last point leads us into a much more convincing response. As I said above, the identity of a platform ought not to be seen as merely a product of the values of its shareholders and core workers, as well as the country in which its main business operations are embedded. Rather, the identity of a platform ought to – under sufficiently inclusive institutional design – be interpreted in Deweyian terms, as heavily influenced by the public (or various publics) interacting with that platform. And if this is right, then I think ultimately the colonialism objection is really an objection to an inadequate process of platform identity formation – one that does not accept input from people in the global South, and hence does not fully permit them to integrate their own reactions to and participation on the platform into their own individual and collective autonomy as equal participants in a social world bidirectionally constructing and constructed by the platforms.

In this way, the developing governance proposals of this book can be partly seen as a (partial) mitigation of the problem of platform colonialism: One of the reasons to accept my argument that we must promote the greater influence of people in the global South on platform governance is that until we do so the platforms will not have a fully integrated identity, capable of administering coherent rules rather than simply operating in a system of, as Ananny and Gillespie (2017) said, "shocks and exceptions." And I think that Schmitt had it exactly backward: Rather than understanding the executive and its power to create the exception as derived from the identity of a people (or a public, as Dewey would say), the identity of a platform public is *developed* over time by the participatory development of institutions that eliminate the necessity of exceptions.

Actually, though, I haven't really drawn this idea from Dewey; it's swiped from James Baldwin:

> The black and white confrontation, whether it be hostile, as in the cities and the labor unions, or with the intention of forming a common front and creating the foundations of a new society, as with the students and the radicals, is obviously crucial, containing the shape of the American future and the only potential of a truly valid American identity. No one knows precisely how identities are forged, but it is safe to say that identities are not invented: an identity would seem to be arrived at by the way in which the person faces and uses his experience. It is a long drawn-out and somewhat bewildering and awkward process.[39]

While Baldwin is talking about the difficult process of creating an integrated American identity through mutual recognition and agonistic reconciliation across its racial divide, a structurally similar point – that the identity of a composite entity or a Deweyian public is learned in part through confronting the demands for inclusion by those who have been excluded – holds in the platform context as well. This argument self-consciously refers to my own prior scholarship, drawing on Baldwin and others (Gowder 2019), about what it might mean for a country like the United States to develop an inclusive constitutional identity over time. At bottom, I think the problem for platforms is similar to the problem for American legal institutions: They were founded on exclusionary terms, and those exclusions have led to claims for inclusion: The only way to avoid persistent lawlessness is with a long-term contestatory process in which those claims are listened to and woven into the pre-existing institutional forms.

A more prosaic way to think about the idea of this chapter is that platform governance institutions are in part developed and functional (or not functional) relative to the identities embedded in the companies as well as the affordances of the platform, as those identities are developed by contestation – including contestation over what governance decisions are made and who gets to participate in them.

Global platforms must have a global identity – the problem of colonialism results, in part, from the attempt to impose a local identity associated with the United States on a global public. But publics are constructed in part out of the institutions through which they might act. And so to genuinely construct a global public, companies must be governed through global democratic institutions. Chapter 6 sketches out one model according to which such institutions might be constructed.

[39] Baldwin (2007, 189); for further analysis of this concept of contestatory political identity in this passage, see Gowder (2019, 397–398).