3

The Problem of Platform Knowledge

Putting a stop to incitement and propaganda aimed at carrying out a genocide is the easiest imaginable moral case for platform governance. The prohibition on genocide is a foundational *jus cogens* norm – a prohibition with as close to universal acceptance as we might want (presumably except by its perpetrators); unlike the more difficult problems discussed briefly in Chapter 2 (such as the Thai lèse-majesté law), there's no basis whatsoever to criticize intervention to prevent genocide as colonial. Moreover, there's a widespread belief among courts, social scientists, human rights scholars, and activists that propaganda like that seen in Myanmar facilitates such genocides (e.g., Wilson 2015; Buerger 2021; Yanagizawa-Drott 2014; Adena et al. 2015; Leader Maynard and Benesch 2016). We can all agree that such propaganda should not be allowed on any platform, and it violates numerous Facebook rules.[1] So what went wrong? The inability of Facebook to deploy local knowledge was the crux of the problem. Facebook lacked reliable ways of finding out about what was happening both on the platform (in the ways to be described below) and in the real world as a result of the on-platform activity.

One incident may be illustrative. In September 2017, "chain messages" were sent over Facebook Messenger to both the Buddhist and the Muslim communities in Myanmar, each warning of impending violence by the other group; according to local organizations, there were "at least three violent incidents" traced to these messages (United Nations Human Rights Council 2018, 341; relying on the account in Phandeeyar et al. 2018). Several civil society groups in Myanmar sent a letter to Mark Zuckerberg relating to that incident which stated that "as far as we know, there are no Burmese speaking Facebook staff to whom Myanmar monitors can directly raise such cases" – and that effectively, Facebook's content moderation system was to wait until civil society organizations sent an English speaker to raise the alarm with the company (Phandeeyar et al. 2018). But the problems ran deeper than local staff: Facebook's software systems weren't set up to

[1] That being said, Facebook saw the need to expand the scope of its rules in order to rectify gaps revealed by the Myanmar genocide (Warofka 2018).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

process the content. Its machine learning systems used to identify things like hate speech couldn't even parse the text encoding in which the incitement was shared (Warofka 2018). While Facebook executives did have *some* warning, as for a number of years beforehand civil society organizations had raised concerns with the company about hate speech in the country (Amnesty International 2022, 51–53), it appears the knowledge didn't lead to action, perhaps because complaints weren't directed to anyone with power, because those making the complaints didn't have the capacity to force the company to pay attention, or because local organizations were unable to convince anyone with authority at the company that genuine physical harms to people were a likely consequence of the propaganda being spread through the platform.

The aftermath of this horror also highlights the fact that the United States is already engaging in colonial platform governance by default. The Gambia appeared in a US court to attempt to subpoena information about the genocidal Myanmar government's communications in order to hold that government accountable at the International Court of Justice. But Facebook opposed the subpoena on the grounds that the Stored Communications Act, 18 U.S.C. § 2702(a), forbade the company from providing the information (Smith 2020; McPherson 2020). In the words of the company, the Stored Communications Act is "a provision of the federal criminal code that protects billions of global internet users from violations of their right to privacy and freedom of expression."[2] (As if the right to privacy and freedom of expression, as interpreted in the United States, applies to a genocidal military on the other side of the world.) Elsewhere in the company's brief, Facebook also makes appeals to a kind of implicit superiority of US law, declaring, for example, that "The Gambia asks this Court to grant it special and unbounded access to account information that no other governmental entity in the world – not even the United States – has available to it in any other proceeding"[3] (as if the level of access of the United States government sets the relevant standard for investigating a genocide), and "Congress's decision to create the CLOUD Act and codify international assistance procedures (e.g., MLAT) reflects its determination of the proper balance between the competing interests at issue in this litigation"[4] (as if the United States Congress gets to decide about the importance of "the competing interests" in a dispute between two totally different countries about a genocide committed by one of them in an international tribunal). Of course, this US-centric rhetoric was necessitated by the fact that The Gambia needed to seek the assistance of a US court to force Facebook to turn over the information, and Facebook evidently saw itself as primarily bound

[2] Facebook's Opposition to Petitioner's Application Pursuant to 28 U.S.C. § 1782 in *In re Application Pursuant to 28 U.S.C. § 1782 of The Republic of The Gambia v. Facebook, Inc.*, case no. 1:20-mc-00036-JEB-DAR (D.D.C., August 4, 2020), p. 1.

[3] Ibid., 9.

[4] Ibid., 9–10.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

by US law (as opposed to, say, that of The Gambia), perhaps because its corporate headquarters and most of its core personnel are located in the United States, which is precisely my point. There is no reason that the people of Myanmar, the people of The Gambia, or the people of the world in general should have to interpret their "privacy and freedom of expression" interests in the ways US law imposes on companies located in its jurisdiction.

One answer to the particular problem faced by Facebook in the Myanmar genocide, had they done so in advance, would have been just to throw money at it: hire more speakers of the local languages as content moderators, task developers to account for minority text encodings. And Facebook surely should have done that.[5] But not even pre-recession Facebook had unlimited resources, and not every knowledge problem can be solved by throwing money at it – one must target one's knowledge-seeking investments, and this chapter is about the organizational background of the conditions under which that targeting can be carried out.

3.1 THE PROBLEM OF KNOWLEDGE: A PERVASIVE CHALLENGE FOR CENTRALIZED GOVERNORS OF DISPERSED POPULATIONS IN CHANGING ENVIRONMENTS

Centralized systems of governance, especially under conditions of diversity, are vexed by the problem of moving knowledge from the peripheries that require regulation to the center from which regulation issues. This problem is the core of leading twentieth-century economist F.A. Hayek's greatest insight: He famously argued that a price system in a free market would be a far more effective regulator of production than any centralized government authority (a la the Soviet Union), because prices capture information about the needs of all buyers and sellers in a market; information that no government agency could ever collect or process. But the problem does not merely occupy the economic domain. Rather, the problem of deploying local knowledge is a key characteristic challenge of states, also highlighted by scholars operating in contexts as diverse as Jane Jacobs's (1992) work on urban planning and James C. Scott's (2008) work on agriculture and central state planning more generally. The commonality among this literature is the insight that the top-down

[5] This is not just a Facebook problem. Whistleblower leaks have identified similar problems at pre-Musk Twitter. In the words of one leaked document ("Current State Assessment," revealed as part of 2022 whistleblower reports and available at https://s3.documentcloud.org/documents/22186781/current_state_assessment.pdf, p. 2), "the lack of diverse backgrounds among employees contributed to gaps in foreign-language and on-the-ground contextual capabilities, hindering Twitter's ability to execute its mission and remove harmful content worldwide. Teams in priority growth markets either do not exist, or are not sufficiently staffed or resourced." As I wrap up the manuscript of this book, it has come out that the major social media companies are still failing at what seems to me to be a much easier case. However badly they're doing at controlling Russian state propaganda about its aggression against Ukraine in English, they're doing even worse in other languages – even in languages spoken widely in the United States like Spanish (Myers and Frenkel 2022).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

viewpoint of a state may be unable to effectively carry out its policies because of its inability to determine the consequences of those policies (or even the needs to be met) in particular local contexts which may be different from the epistemic environment of the policymakers.

I shall suggest that those insights apply just as well to private governance – actually more so. First, however, we should delve a little bit more into the intellectual background to get a fuller sense of the commonalities among the settings of and proposed solutions to the problem of knowledge.

Jane Jacobs's *The Death and Life of Great American Cities* turned the profession of urban planning on its head in 1961.[6] In a reaction against centralized models of utopian urban planning associated with architects and planners of the likes of Olmstead, Moses, and Le Corbusier, Jacobs understood urban planning (and larger economic processes – see discussion in Desrochers and Hospers 2007) as fundamentally organic, bottom-up, improvisational, and adaptive. Centralized planning – the gigantic housing project, the "slum clearance" – in such a context could be actively value-destroying, because existing urban forms of organization, even if they looked messy, represented successful adaptations to the challenges distinctively faced by the people of a place in the context of their own social, economic, and physical lives – in Jacobs's (2016, 75) words, they incorporated a kind of "weird wisdom" derived from the activities of the people occupying that space.

For Hayek, it would be easy for a policymaker to be tempted by Soviet-style central planning. For the central planner, it might seem sensible to make a decision like "we need this many tires, this many gallons of oil," and to governmentally control production in order to generate those goods. Our planner thus avoids all the complexities of free markets, such as inequalities of wealth (and hence of access to human needs), wasteful production of unnecessary luxuries, redundancies in suppliers, and so forth. But in identifying the trap into which our planner has fallen, Hayek (1945, 524–26) too focused on dispersed knowledge and on adaptation: It turns out that it's actually impossible for central planners to know things like how many tires are needed, because that's just the aggregate of the needs of everyone else, which are hard to discover; moreover, even if the central planner could know how many tires were needed at a specific point, it would be impossible to adapt to changes in local conditions which might shift where and how many tires are needed in an instant. However, free markets – and in particular the price system of such markets – could effectively aggregate that information, because people respond to local changes in needs by changing their willingness to pay. If tires are particularly needed in a specific community, they will out-bid other communities for those tires, and thereby communicate their need to producers – with accompanying incentives to take that information into account – without any of the participants needing to know anything other than the change in the prevailing price.

[6] See discussion in Campanella (2011).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

Observe how Jacobs and Hayek share an emphasis on a key problem that applies in the platform context with some salience: *adapting pre-existing plans to change, where information about that change is spread out across the governed territory*. Moreover, the key method of such adaptation is to give those with local knowledge about conditions and needs some direct *control* over outcomes. Thus, a planner who gets the answer to an immediate problem right and makes commitments – whether those commitments are the resources to be allocated to particular production and distribution networks or the buildings to be built – cannot adapt those commitments to exogenous shocks unless people with both the knowledge and the incentives to respond to those shocks are capable of exercising some control over them. Both Hayekian price systems – which encourage producers to change pre-existing resource commitments in pursuit of profit – and the organic development of urban landscapes – which permit relatively fast-paced and small-scale adaptation – can serve those adaptive functions.

Some decades later, James C. Scott (2008) integrated the – by then widely accepted – insights of both Jacobs and Hayek into a broader analysis of the failures of a style of centralized government intervention that he called "high modernism." In a way, on Scott's account, centralized authorities recognize that they have a problem with learning the information necessary for their projects to succeed – and their response is to attempt to impose "legibility" on the world. The idea of legibility effectively highlights a core idea shared by Hayek and Jacobs, namely that the distinctive informational problems with which they wrestle are the product of complexity: Cities, like economies (and, I submit, like platforms) feature dense interactions among numerous and diverse actors, exponentially increasing the informational challenges associated with centralized planning. Thus, for Scott, the response of state bureaucracies is to attempt to reduce the dimensionality of the problems with which planners are presented: to summarize them only in terms of the particular issues with which they are concerned as matters of policy and then, in effect, attempt to optimize along those lines (compare Reich, Sahami, and Weinstein 2021 on pathologies resulting from narrowly focused technology company optimization). But such a myopic approach tends to lead to disaster, as the needs and values as well as the local knowledge and adaptive capacity of those whose lives are thereby reorganized tend to be disregarded. Thus, Scott describes the failure of a wide variety of high-modernist planning projects, ranging from Soviet and colonial agricultural "reforms" to (in passing) Taylorist factory management to the creation of planned cities like Brasilia.

In the abstract, we might summarize the core lesson of Jacobs, Hayek, and Scott as a pattern for the solution of these kinds of knowledge problems. When important knowledge is in the periphery and cannot be effectively directly moved to the heads of central decision makers, individuals must first be given an incentive to use their knowledge – that is, doing so must produce useful outcomes for them as individuals. Those uses of knowledge must also, in the aggregate, be capable of correctly influencing aggregate outcomes. Hayek's account of prices, thus, accounts for both the individual incentives – people need to use their knowledge of their own

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

needs and resources in offering and demanding goods in markets – and the aggregate influence – the overall prices (under the right conditions) successfully reflect the conjunction of needs and resources of all in the market. Similarly, the existing urban structures in Jacobs's weird wisdom represent the successful use of individuals' knowledge to solve individual problems but are (quite literally) built into urban structures that continue to be functional and to interact productively with similar uses of knowledge from other individuals.

Many of the problems of platform governance can be understood as center-periphery knowledge problems of the sort considered by Jacobs, Hayek, and Scott.[7] Indeed, based on what we know about the class of problems, we ought to have expected them to be the characteristic challenge of platform governance. For such problems are rooted in scale and diversity: As the number of people the governors have to regulate grows and as the social distance between the governors and the governed grows, it becomes more difficult for those doing the regulating to have any idea either about the needs or the activities of those regulated. Add in the capacity of platforms to create network configurations both dense and dispersed in ways previously unavailable to human sociality, and the emergent patterns of behavior that we've all watched result from that capacity over the last few years – as well as the consequent high rate of innovation, novelty, and chance – and it should be unsurprising that platforms have had persistent problems with learning about user behavior and interests as well as adapting to changing behavioral environments.

Thus, consider as a salient example the various missteps of over-governance on social media – circumstances where social media companies have misinterpreted subcultural communication or failed to attend to the local context, and, in doing so, created scandals over their removal of socially beneficial content. Just on Facebook alone, for example, there have been famous scandals involving the removal of famous Vietnam War Photos (Scott and Isaac 2016), requiring transgender people to use their birth (dead) names (Holpuch 2015; see discussion in Haimson and Hoffmann 2016), removing as hate speech content of LGBT users who are reclaiming terms like "dyke" and "tranny" (Lux 2017), and treating topless photos of (and posted by) indigenous women in traditional settings as forms of obscenity (Alexander 2016). Nor are other companies free from similar problems; for example, researchers found that Google-developed artificial intelligence to detect online "toxicity" disproportionately identified the speech of drag queens as toxic (Dias Oliva, Antonialli, and Gomes 2021). Even as recently as 2021, these problems persist, even in Burmese (the one language you'd think Facebook would

[7] Frank Pasquale (2018) has a particularly interesting take on platform knowledge. He begins with the notion that platforms are in some sense a *solution* to the problem of knowledge, insofar as their immense amount of data and machine learning capacity permits them to organize far more activity than the state when Jacobs, Hayek, or even Scott were writing. But this turns out to be illusory, for platforms too struggle to manage such a vast scope of activity, and they end up becoming "absentee landlords" when they cannot manage their gigantic domains (or choose not to care because neglect is profitable).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

try to get right after the genocide): When Burmese speakers advocated borrowing financial strategies from Hong Kong's resistance to the Chinese Communist Party to resist Myanmar's authoritarian government, Facebook misinterpreted it as hate speech against Chinese people, necessitating a reversal by the Oversight Board.[8]

Moreover, as James C. Scott would predict, efforts to address the problem of knowledge by rendering human behavior on platforms legible can distort governance choices. For example, Siapera and Viejo-Otero (2021) describe how Facebook's efforts to address hate speech have manifested as the kind of "colorblind" (and gender-blind, disability-blind, colonialism-blind, etc.) rules critical race theorists have aptly criticized in US law for failing to take account of underlying social hierarchy in interpreting hierarchy-laden concepts like racism (e.g., Gotanda 1991). Although Siapera and Viejo-Otero read this colorblindness through an ideological lens, they also discuss what seems to me to be a more primary cause, if not of the adoption of colorblind policies then certainly of their resilience against external critique: The difficulty of identifying the relevant relationship between social hierarchies and language in different societies, that is, a problem of legibility. This is illustrated by an extended quote from Mark Zuckerberg, which leads Siapera and Viejo-Otero's article and which attributes Facebook's colorblindness to the problem of dealing with vast diversity in social meaning at scale:

> [W]e don't think that we should be in the business of assessing which group has been disadvantaged or oppressed, if for no other reason than that it can vary very differently from country to country […] It's just that there's one thing to try to have policies that are principled. It's another to execute this consistently with a low error rate, when you have 100 hundred billion pieces of content through our systems every day, and tens of thousands of people around the world executing this in more than 150 different languages, and a lot of different countries that have different traditions.[9]

3.2 THE DEMOCRATIC, DECENTRALIZED SOLUTION TO THE KNOWLEDGE PROBLEM

Broadly speaking, we can think of Hayek and Jacobs as offering solutions to problems of knowledge rooted in individual agency – individual decisions of buying and selling in a market which aggregates those decisions for Hayek; individual decisions in land use aggregated by the fact that those decisions are also influenced by the nearby individual decisions of other land-users for Jacobs. But these are modes of knowledge aggregation that most plausibly occupy what we might think of as *productive* contexts, as opposed to *regulatory* contexts – that is, circumstances when

[8] Content Moderation Oversight Board, decision in case FB-ZWQUPZLZ, https://oversightboard.com/decision/FB-ZWQUPZLZ/.

[9] Siapera and Viejo-Otero (2021), 113, quoting leaked statement from Mark Zuckerberg.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

people are building or consuming things, and they have strong incentives to make the things they build or consume compatible with one another. It is less obvious how such patterns might be adapted to a purely *regulatory* context like platform governance. In conventional polities, for example, we do not let people evolve bottom-up rules of criminal law from their individual choices.[10] Instead, a different structure of aggregating individual knowledge is needed, and the standard approach is to connect democratic political decisions to collective learning. Dewey (1927, 206–8), for example, argued against expert government on the basis that democratic processes permit ordinary people to communicate their knowledge and needs upstream.

The abstract logic of this argument is tempting: Democratic institutions give people an incentive to contribute their knowledge – their votes may produce outcomes that are to their benefit – and translate those uses of knowledge into collective outcomes. Democracies start to sound much like Hayekian markets. But – in part because this idea comes so naturally to those of us steeped in the ideological environment of contemporary western liberal democracy – the democratic solution has been vexed by a number of important challenges to both ends of the solution pattern. On the collective outcomes end, a key challenge has come from social choice theory, which captures the idea that it's surprisingly difficult to translate individual preferences (hopefully reflecting knowledge) into aggregate outcomes (e.g., Riker 1982). On the actual knowledge end, it's widely recognized that democratic institutions tend to give people votes on things about which they don't necessarily have particularly meaningful knowledge, and, moreover, may give individuals insufficient incentive (due to the low probability that their votes may deliver any desired outcome) to acquire or use relevant knowledge (e.g., Achen and Bartels 2016; Brennan 2016).

And yet there is also an important literature on the capacity of democratic institutions – if the incentives are right (e.g., if there are shared preferences as to common goals) – to structure the effective aggregation of knowledge in ways that at least potentially avoid these problems. Particularly important is Ober's (2008) account of how the network structure of Cleisthenes' reforms in sixth-century Athens helped the *demos* effectively leverage dispersed knowledge (creating, I daresay, the very first networked leviathan). On Ober's account, Cleisthenes' system combined strong-tie network structures in local demes (essentially, neighborhoods/villages) which permitted people to develop dense social knowledge about the knowledge, skills, and character of their neighbors, with artificial forms of higher-level social structure that permitted the formation of weak network ties between demes. These included "tribes," which were administrative and political units which were created to combine demes from coastal, inland, and urban regions (hence building geographically dispersed network ties), and the Council of 500, the main executive body of the city, which was made up of randomly selected representatives from each tribe.

[10] I confess that we might tell a story about American criminal law's emergence from English common law which interprets them that way.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

Ober teaches us that the network structure of democratic decision-making matters. In the abstract, he offers a case for a structure that promises people influence (either locally or by integrating them into centralized decision-making mechanisms in such a way as to give them a visible impact on decisions at least some of the time, as by sortition-based councils), thus giving them an incentive to put their knowledge to use; as well as broader and socially novel network connections, to transmit knowledge among individuals and to socially distant decision makers.

Note how Ober's (2008, chapter 4) account of Athenian knowledge aggregation mechanisms describes key functions to handle the problem of aggregating, centralizing, and retaining diverse knowledge. The Council of 500 aggregated widely dispersed knowledge from all sectors of society, and operated essentially continuously while giving members incentives to build novel network interconnections with unfamiliar persons. Individuals only had short terms on the council and were likely to receive reputational rewards for successful leadership. They had strong incentives to interact with people who would otherwise be socially distant from them in order to succeed in their Council roles. After their terms were up, they were also likely to reap economic benefits from the distant connections that council service made available. Moreover, because of the density of council service within the community (due to short terms and sortition), that knowledge could be further dispersed as well as transmitted to new members.

What might all of this mean in practice in the modern context? Specific proposals will be left to Chapter 6, but Athens-like structural innovations seem tailor-made for handling platform problems involving dispersed knowledge. Imagine how the Myanmar disaster could have gone differently if a local (and socially dense, in a network sense) group of Burmese citizens and activists capable of swiftly and credibly identifying the presence of military propaganda on Facebook was directly incorporated into company decision-making processes via some similar kind of council and had already built relationships, via those processes, both with corporate personnel and with similar local groups elsewhere? Suppose that Facebook's internal bureaucracy was already accountable to such a council, which was diverse but included ordinary Burmese people, who regularly met to consider updates to platform rules and enforcement priorities? Such a council could have escalated the risk of genocide to the company earlier – and would have had an incentive to do so, if its word actually counted in company decision-making (i.e., if the participation of ordinary people was efficacious).[11] Or, perhaps, it could have escalated the

[11] Had Facebook been alerted of the danger, it's likely that the company could have at least done something relatively quickly, such as a patch to their content moderation tools to handle the text encoding in which the hate speech was rendered at a minimum. Some suggestion that there was low-hanging fruit available to at least mitigate the risk comes from the fact that after the scandal, the company supplemented its automated content moderation and managed to begin to do a substantially better job in Myanmar (Gorwa, Binns, and Katzenbach 2020, 2).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

problem to international human rights agencies. Put a different way, it could have leveraged local knowledge to more readily identify community needs.[12]

Note how this is distinct from the task of merely hiring more content moderators. In the platform economy, as it currently stands, content moderators are substantially disconnected from actual company decision-making processes: As Sarah Roberts (2019) describes, such moderators are typically hired by contractors, physically isolated from company decision makers, and managed in a high-rate task-oriented fashion that gives them little opportunity either to deliberate about the problems they see or communicate with decision makers about priorities.

Another way of putting the problem is in Scott's terms. In the first instance, what matters to Facebook is the match between content on the platform and the rules they've written down; the job of content moderators is to make that content legible in those terms. Accordingly, the sorts of knowledge which content moderators may leverage are chosen in advance by the company: They may use their knowledge of the local language and idioms to identify whether some particular platform activity violates a closed set of rules. They are not asked to use their knowledge about the overall conditions and needs of the communities of which they are members. Even moderators who have worked on the main company campuses as contractors have reported that their input is not welcome (Roberts 2019, 97).

That knowledge is, from the perspective of platform governance, simply wasted – but it need not be wasted and could be deployed with the creation of democratic decision-making processes accountable to people in those communities. Indeed, this could include the content moderators themselves along with ordinary users, to the extent they could win or be granted (e.g., by states) labor rights entitling them to some degree of participatory decision-making in their workplaces. More empowered content moderators with a real voice in their workplaces and real career paths would be in better positions to communicate knowledge that they have to company policymakers.[13]

Chapter 6 will propose the creation of randomly selected grassroots governance organizations across the world as the primary method to integrate peripheral knowledge into platform companies. Hélène Landemore, another key democratic theorist of knowledge, has argued (Landemore 2013, 2020), based on the analytic work of scholars of diversity and democracy such as Hong and Page, that sortition

[12] Compare Anderson's (2006, 17–21) discussion of Bina Agarwal's (2000; 2001) work on forestry decision-making in South Asia, which offers a Deweyian interpretation of Agarwal's research on the way that the exclusion of women from many local groups has led to failed decision-making because that exclusion also sacrificed knowledge about things like how to enforce resource management rules and how much wood may be safely gathered.

[13] Organizational sociologists have noted that personnel moving between roles is a key mechanism of organizational and intraorganizational learning and may also contribute to "redundancy in mechanisms for sensing and responding to change" Westley (1995) 416–417. Making content moderation less of a dead-end job and more of a job that could rotate into company decision-making roles as well as jobs with regulators and the like has additional potential to more fully leverage local knowledge in this sense.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

is epistemically superior to other methods of democratic selection in virtue of its capacity to generate diverse pools of decision makers. (Classical Athens, not incidentally, is famous for its use of sortition to fill decision-making bodies.)

Sortition is likely to be a key selection process for any democratic form of platform governance for several additional reasons in addition to its positive effects on the ability of such processes to leverage knowledge from persons socially distant from platform employees. First, in the context of extreme global institutional diversity, sortition seems likely to be much easier to administer than other selection methods, as it would not require adaptation to specific versions of more active political selection (such as campaigning and elections) with which users may be disparately familiar. Relatedly, sortition at least partly routes around institutional corruption and subversion – to the extent any other method of selection requires active administration of things like voting rules by either companies or governments, sortition reduces the risk that such administrative processes can be used to put a thumb on the scale.

That being said, it is unlikely that entirely sortition-based democratic processes will be sufficient. There is also strong reason to guarantee (and not just probabilistically) representation from a wide array of discrete social groups in platform decision-making. Moreover, sortition would be most practicable to administer from among platform users, but nonusers and groups predominantly composed of nonusers are also the victims of platform externalities and are likely to have helpful knowledge to contribute about it (as well as moral claims to participation in its governance). Nonetheless, the advantages of sortition are sufficiently compelling that it will certainly be a key component of any effective democratic platform governance system.

Another knowledge-related advantage of the creation of broad-based participatory democratic institutions flows in the other direction, that is, from platforms to the public. As Cohen (2019, 181) has aptly argued, regulatory strategies for things like the abuse of platforms that rely on transparency are unlikely to be successful, because under conditions of "infoglut," the challenge is not accessing information but interpreting it. (Have any of the readers of this book ever looked at the ad transparency database Facebook rolled out after the 2016 disasters?) But one solution to infoglut is specialization, that is, to delegate to a part of the global public at large primary responsibility to process such information. Traditionally, such roles are held by researchers and journalists, but both are also subject to infoglut,[14] as well as to other institutional pressures as well as a lack of direct access to relevant decision makers. Participatory democratic platform governance institutions could in principle be given such direct access, for example, to regulatory officials of the states in

[14] Go ahead, just ask me about how easy it is to turn in a book manuscript within a year of the deadline written in the contract about something as timely as platform governance, about which there is a constant flood of new scholarly publications as well as evolving facts on the ground. In unrelated news, my most sincere apologies to my long-suffering editor at Cambridge and grant administrator at the Knight Foundation.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

which platforms operate, and may benefit from sufficient specialization over a short term to be able to focus on and process information coming from platforms, avoiding the problem of infoglut. (We might compare this to the specialized magistrates in Athens who took time off from their ordinary lives to manage particular areas of Athenian government.)

In addition to giving people an incentive to contribute their knowledge,[15] participatory governance can also bring together knowledge and legitimacy. The sociological concept of the legitimacy of a system of governance refers to the notion that people will actually accept governance arrangements and hence will comply with the outcomes of those arrangements voluntarily rather than by force (Tyler 2006). One common proposition in the literature on sociological legitimacy is that voice or participation is a source of legitimacy in this sense (Tyler 2009, 319).[16] There is also evidence that people given an opportunity to participate in jury-like resolutions of social media content moderation cases perceive those processes as more legitimate, although not necessarily more effective (Fan and Zhang 2020).

We might, however, helpfully disaggregate the relationship between participation/voice and legitimacy into two plausible causal directions. First is the conventional supposition that to the extent people are given an opportunity to participate in rulemaking and application/enforcement processes, they will perceive the decisions made under those rules to be more legitimate, and hence be more likely to comply with them. The second, perhaps more interesting, idea reverses the causal direction and observes that participating in collective decision-making is costly (as any reader who has ever been in a faculty hiring meeting can attest), and hence that recruiting people to participate requires giving them some reason to think that the ultimate decisions will be something they can accept – for example, by giving some guarantee that their participation will be effective at preserving their core interests, that is, providing avenues for what we might (mildly tendentiously) call *real* or *meaningful* participation. This reverses the causal relationship in the sense that it derives participation from legitimacy – it suggests that the creation of opportunities for *meaningful* participation, opportunities that respect the interests and status of potential participants, might help give those thus respected a reason to participate.

Both of these causal directions are at play in the claim of this chapter that dispersed and distributed governance, done right, has the capacity to improve effectiveness in two ways: by recruiting the knowledge of participants and by recruiting their voluntary compliance (and hence reducing monitoring and enforcement costs). But the *meaningful* participation proviso has other important benefits, for sometimes people who only enjoy nonmeaningful participation are simply ignored

[15] Cf. Fung and Wright's (2003, 25) related concept of "empowered participatory governance."

[16] There is some question about the empirical robustness of this proposition (Hibbing and Theiss-Morse 2008). Still, there is some evidence from the new-governance-style networked governance arrangements on which this volume partly draws that greater degrees of voice are associated with greater degrees of perceived effectiveness (Klijn and Edelenbos 2013).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

when they try to contribute their knowledge. This too happened in Myanmar with Facebook: Civil society organizations tried to tell Facebook what was going on but were ignored (Amnesty International 2022, 51–53). Had those organizations or ordinary Burmese people the power to directly intervene on content moderation decisions in the country, they likely could have gotten more serious attention from more powerful company personnel, if only because appeasing them would become more necessary in order to meet more of the company's overall goals.

3.3 PARTICIPATORY GOVERNANCE FACILITATES LEGIBILITY

It will be useful, to fill out the character of the knowledge that ordinary people as well as presently disempowered workers might introduce in some more detail, to disaggregate the problem of local knowledge into distinct problems of *observability* and *legibility*. This is not meant to be a robust typology but rather a pair of ad hoc constructs (partly derived from Scott) to help us distinguish the ways that governments might be challenged to know about behavior and interests and the ways that platforms face similar challenges.

Legibility as a concept has been used prominently in studies of the institutional design of state authority (Scott 2008) as well as urban design (e.g., Lynch 1960). For present purposes, we can think of legibility as referring to the capacity of an observer, having observed some behavior, to classify it in some pre-existing conceptual scheme. The concept of legibility is important in the nongovernance platform context, because, as Julie Cohen (2019, 37–40) explains, providing legibility of persons to advertisers is a key economic function of platforms. In the present work, however, I focus on the problem of platforms acquiring legibility over user behavior.

What might it mean for a platform to "know" what some behavior "is?" Typically, platforms have a high capacity to observe conduct on their platforms. After all, a key distinction between internet platform companies and every other form of governance is that every interaction within the governance scope of platform companies is hosted on infrastructure in the control of the platform; with the exception of end-to-end encrypted communications (most importantly, WhatsApp) and certain kinds of decentralized technologies such as the Mastodon federation,[17] the companies under discussion here have the power to inspect the technical contents – the ones and zeroes – of any byte flowing through their network assets – any financial transaction, any Instagram image is visible in that limited sense.[18] This is in contrast

[17] As a decentralized network of platforms, the Mastodon federation has the potential to upset many of the suppositions on which the platform economy runs. However, the technology is only coming into relatively widespread use after the Musk takeover, and hence has not been observed sufficiently long for me to make any confident statements about it; moreover, it may not support for-profit financial models.

[18] Blockchain and other kinds of decentralization technologies may change this; however, I lack the expertise – or, quite honestly, the interest – to consider blockchain; it is also very difficult at best to come to any plausible guesses as to the likely impact of blockchain on all the business areas currently

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

to governments, which must expend resources on surveillance to observe any particular thing happening in their domains of governance.

For a concrete example: As I type the words of this chapter, chances are that the government has no clue that I am doing so. As far as I know, there are no government cameras in my office; no prosecutor or intelligence agency has gotten a warrant to search the contents of my computer. Maybe the NSA is hoovering up the data that my machine sends over the network, but it doesn't directly control the software on my computer and I'm not remotely important enough to warrant the attention of serious government hackers, codebreakers, exploiters of zero-days, and so forth, so, as long as the various encryption algorithms that my computer automatically uses to send my writing around hold, I can feel pretty confident that the government won't "know" this paragraph exists, in the extended sense of "know" that means "some government official can look at it without getting a warrant to intrude on someone else's property rights by force" at least until after the book is published.

By contrast, at least one platform company can look at it at will. I store my working drafts in markdown files in private repositories on GitHub, which is owned by Microsoft. For all practical purposes, Microsoft can observe this text shortly after I write it. And, in all likelihood, were I to include some content in this draft that violated an important legal obligation either to Microsoft or of Microsoft – for example, if I uploaded the Windows source code to the repository for this book, or child pornography – automated systems would probably detect that and take the content down and/or alert Microsoft personnel long before the government could find out.[19]

However, they may lack the capacity to *interpret* content on their platforms, that is, either to translate bits into communication (as with the task of text and photo identification), or to discern the meaning of communication (as with linguistic translation as well as subcultural message interpretation). This is the problem that

encompassed by the platform economy. The technology has been around for over a decade now, and doesn't yet seem to have had much impact beyond creating a bunch of extraordinarily energy-hungry speculative bubbles. (Also, I think, albeit with very low confidence, that blockchain actually would increase overall observability of some activities to the extent it replicates records of those activities in a distributed database?) As for end-to-end encryption, to my mind this is a feature of platforms that they themselves control – Meta doesn't have to offer end-to-end encryption on WhatsApp; its executives choose to do so. In principle, if they have a sufficient interest in controlling conduct carried out through that platform the company could cease encrypting it. Alternatively, some kinds of governance could in principle be pushed to the application layer – for example, by building certain machine learning classifiers for content like hate speech directly into client applications (which see cleartext) rather than the network, subject to available computing resources on client devices (which may be a hard constraint in some particularly sensitive markets for services like WhatsApp, such as India).

[19] That's speculation, but it's pretty plausible speculation: I mentioned child pornography as one of my examples because Microsoft has actually invented one of the leading technologies to detect and block known child pornography images; given the extremely harmful and criminal nature of such content, it would be surprising indeed if they didn't have the sense to use it on Github. See Microsoft, PhotoDNA, www.microsoft.com/en-us/photodna (visited April 3, 2021). (Microsoft describes some of its own use of the technology on its services in Langston 2018.)

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

I call “legibility” and it should by now be obvious that I do so because it’s useful to swipe some ideas from James C. Scott. Interpretation depends on context: External knowledge can make a given observation more or less legible.[20] The interpretation of idioms is one example: Speakers of British and American English typically have different knowledge about the meaning of sentences including local slang; one person’s homophobic slur is another’s cigarette. But noncommunicative intentions can also be the subject of interpretation. For example, Facebook engages in efforts to police “coordinated inauthentic behavior” where different accounts work together to deceive users (Gleicher 2018). Such coordination is often facilitated by offline interactions which Facebook cannot directly observe, with the quintessential example, of course, being the “Internet Research Agency,” a Russian intelligence operation to spread political disinformation across social media. Interpreting a given piece of content as genuine political advocacy, as opposed to coordinated inauthentic behavior, depends on extrinsic knowledge about, for example, the identity of the speaker and their other social affiliations.

Intuitively, and without delving into the philosophy of language, we can fairly confidently assert that legibility comes in layers, where understanding of a layer depends on the external knowledge one brings to an interpretation. A stream of bits passing through some router might be legible as text rather than an image if one can interpret the metadata attached to its network packets. With a little more knowledge (about which text encoding is used), the text can be deciphered into letters. With still more knowledge, those letters can be assembled into words in a known language. The meaning of those words further requires social context: What a given set of words in a particular language means to speakers and listeners in one social milieu may be completely different from those in another. The obvious concrete example of that last stage with which most readers will be familiar is the deep complexity surrounding the use of certain words which have traditionally been race or gender-based slurs: Used from a member of an advantaged group to a member of a historically oppressed group, such slurs can be among the vilest of hate speech; the same words can also be reclaimed and used within historically oppressed groups as a signal of affinity and solidarity.

For an observer, legibility is a classification exercise relative to a particular task.[21] Seeing a stream of bytes passing through a network, an observer interested in knowing whether it is pornography and one interested in knowing whether

[20] For example, PhotoDNA renders raw bits of image data more legible as child pornography by matching them against context, where that context is the universe of images previously identified by human observers.

[21] This is also true of the other sort of legibility in which platforms specialize, namely the legibility of users to advertisers. As Cohen (2019, 71) notes, the purpose of that kind of platform legibility “is not understanding but rather predictability in pursuit of profit”; in other words, an advertiser does not care whether a platform’s profile of an individual is accurate, it cares whether a platform’s model can accurately predict which users are likely to buy their products.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

it is political dissent may enjoy different levels of legibility, depending on their capacity to understand the ways in which the sender and the receivers of the bytes express their sexuality or their politics. This is Scott's (2008, 22–23, 80) point in deploying the concept: Raw unmediated data from the world isn't usable in any project of governance; rather, that data must be slotted into administratively meaningful categories; to translate from some-stuff-happening-out-there to "this is political dissent," "this is porn" is to make it legible. For that reason, the interpretation of our social world is typically a goal-directed activity. In fact, depending on the interests of an observer, the same image might be pornography *and* activism – for example, a sex workers' rights activist might make activist pornography, which may be interpreted just as pornography (and hence a fit subject for regulation) by a platform employee tasked to enforce rules protecting children from sexual content, and just as activism by a platform employee or user interested in promoting debate on sexual freedom.

A point that I have not made explicitly but which should be apparent from the last two paragraphs is that many of the most important kinds of legibility at play in both governments and platforms depend on inferences about the cognitions that actors, speakers, and listeners have in their heads. This is not to take a position on some kind of deep question of linguistic meaning – I don't care whether or not the author is dead – but simply because the *practical purposes* of those who try to interpret communication often involve drawing inferences about the intentions of speakers and actors in addition to things like the effects of such speech on listeners and observers. Thus, in order to figure out whether some transaction carried out on a platform is actually an illegal arms or drugs sale, we need to know whether the messages that are sent over that platform constitute a shared intention to cause one participant to send some weapons or drugs to the other. In many cases, the contextual knowledge required bears on the pre-existing set of assumptions and understandings shared by the users of a platform which allows them to form these shared intentions. This is why the hate speech problem is so difficult: Hate speech is deeply contextual; the n-word (and related spellings and pronunciations) has very different meanings depending on a wide variety of facts about who is addressing whom with the term, where even the most obviously relevant facts (i.e., the racial identities of the people involved) are themselves complex and context dependent (Gowder 2015).

Because of the vast cultural diversity on platforms, the contextual knowledge necessary to attain high levels of legibility is likely to be particularly difficult for platform companies relative even to governments. Returning to the example of the private GitHub repository that contains the draft of this book: Microsoft is likely to be able to identify the contents of this repository as English text – it's in a standard UTF-8 text encoding; machine language recognition is easily sophisticated enough to identify the English language from a sample of hundreds of thousands of words – and in view of the subject matter, it's likely that most any Microsoft employee who looks

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

at it will have a general idea of the content. That would not likely be the case if this were instead a book in, say, Tamil, and if the subject matter were, say, the influence of Immanuel Kant on evidentiary approaches in eighteenth-century Bulgaria. However, even with respect to this book, Microsoft would be obliged to spend some resources to reach a very advanced level of legibility – the book makes references to scholarly conversations which some, but not all, Microsoft employees would have the background to interpret without doing additional work.

Wilson and Land (2021, 1060–69) describe the problem of platform regulation of hate speech in similar terms. On their account, a key issue is that such regulation lacks sufficient "context" – where that context derives from an understanding of the underlying community. Even if a company can identify something like hate speech in the abstract – saying nasty things about an ethnic group, for example – social context is required to understand the difference between, say, a group that is socially subordinated complaining about that subordination and incitement by a dominant group to violence. Moreover, the practical impact of that speech, as they note, may differ according to local context about which moderators may be unaware. In their words:

> A content moderator located in a distant country may not be aware of widespread election unrest, outbursts of communal violence, or a pattern of violence against sexual minorities in a locale. Since each moderator only sees a small sliver of the total range of expression about a topic or person, campaigns of systemic harassment are harder to identify.[22]

Even though, as Scott detailed at length, legibility is a core problem for governments, nonetheless, for most kinds of communication, the greater linguistic and cultural proximity between governments and their people is likely to facilitate legibility, relative to platforms.[23]

Now return to Facebook in Myanmar. Unlike a local government, Facebook would not necessarily – and at the critical points when it should have denied its services to the perpetrators of genocide did not – have personnel with linguistic and cultural competence to interpret communications over its network assets in Myanmar. Accordingly, where a local government might be able, could it observe certain messages transmitted over Facebook assets, to understand what was being said, Facebook could observe such messages at will but not interpret them.

Of course, in the case of the Myanmar genocide, the government couldn't have helped, as it was a perpetrator. But if we imagine a counterfactual nongenocidal

[22] Wilson and Land (2021, 1067).

[23] The classic example of the gains to legibility from cultural proximity comes, of course, from the famous Navajo code-talkers of World War II: because the United States had a much easier time than Germany or Japan in finding and hiring people from the Navajo Nation, the cost to make communications in code in Navajo legible was much, much lower for Americans. Apparently a key advantage of the code-talker system was the distance between the sounds produced in Japanese language and those produced in Navajo, which stymied even consistent transcription of the coded communications (Huffman 2000).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

government trying to address private speech that threatened ethnic violence against the Rohingya people, the contrast between that government and the platform couldn't be clearer: Facebook had the capacity to observe every message sent through its servers but couldn't actually figure out what it meant without third party assistance or costly (and all-too-late) investments in native speakers of the language who understood the culture (and technical investments in parsing the text encoding called Zawgyi). The (counterfactually nongenocidal) government was filled with native speakers who understood the culture, and so in principle could understand most messages transmitted over the platform, but had no practical way to observe those messages in the first place.

However, one of the key lessons from scholars like Scott and Jacobs is that even achieving legibility in the limited sense of knowing at a sufficient level of detail what some language means is insufficient. The discussion thus far is still within the realm of hiring more content moderators, at greater or lesser degrees of individual empowerment – to interpret individual pieces of content in light, that is, of the ability to use greater or lesser degrees of local knowledge about the meaning of that content. But some sorts of legibility are in some sense only accessible in the context of a design that is itself derived from local knowledge rather than centralized command. This is part of the point of the theoretical work on democratic knowledge I described a few paragraphs ago: Another way of reading Ober's insight is that democracy creates feedback loops that can make use of local knowledge. A centralized decision which, for example, creates rules exploitable by a genocidal government cannot be rectified just by taking down more and more hate speech; rather, that decision itself needs to be subject to *feedback* based on knowledge about things like the actual local threat of genocide, which can then be routed back into the design of rules and enforcement systems sensitive to the local context.

We can borrow from one more participatory context to learn about the usefulness of participation for legibility, to wit, the common law. In American law, contextual knowledge (or at least judgment) is often specifically incorporated into the criteria for triggering legal consequences. For example, in articulating the test for whether some speech is outside of the protections of the First Amendment on grounds of obscenity, the Supreme Court has declared that an obscene work must be one as to which "the average person, applying contemporary community standards would find that the work, taken as a whole, appeals to the prurient interest."[24] This reference to "contemporary community standards" directly invokes the contextual knowledge of some (not terribly well-specified) community external to the legal system whose standards are to be used. Consider also the appeal to the "reasonable person" in setting the standard of care in negligence law.[25] In general, the common law frequently appeals to community understanding captured in various reasonableness doctrines.

[24] *Miller v. California*, 413 U.S. 15, 24 (1973).

[25] See discussion in Gilles (2001).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

It's no coincidence, I think, that common law systems such as the United States and England also feature the substantial use of juries. There's a kind of inherent connection between those two systems: Part of the underlying theory of the common law has been that it is genuinely common, that is, that it derives from the experience and values of the community in organizing their lives.[26] And sole elite judges are naturally less competent in making that judgment than a diverse group of ordinary people from the community, in a deliberative setting in which they can draw on and reconcile their own experiences of community norms.

There is also a close connection between the form of the jury and the resistance to the oppression of minorities, as the demand to be included in a jury can amount to a demand to have one's own knowledge and judgment included in the process of applying the rules used against one. The famous Ten-Point Program of the Black Panthers expressed the point most vividly: because Black people were excluded from juries, "We have been, and are being, tried by all-white juries that have no understanding of the 'average reasoning man' of the black community" (Bloom and Martin, Jr. 2016). In other words, one of the ways that unjust hierarchy could be implemented in a legal system – especially in a common-law-derived system such as that of the United States – was to exclude the local knowledge and reasoning processes characteristic of subordinated groups from the adjudication of cases involving them.

This is a point that directly relates to the problem of platform colonialism discussed in Chapter 2. Legibility is a moral as well as a practical problem. When Twitter decides that an exception for enforcement action will be made for "legitimate public interest,"[27] the concept itself assumes a reference to an identified public whose interests are at stake that might be wholly distinct from the people in San Francisco who ultimately make the decision. In effect, in an enforcement content, "legitimate public interest" functions sort of like "reasonable person" in the law of negligence, that is, a reference to a community whose judgments the decisional process is meant to replicate. And in the context of global inequality, in which many people in other countries rightly perceive that Americans do not understand or respect their political processes and cultural values, the need to make judgments about things like "legitimate public interest" generates a demand to include the public whose interest is being legitimated (or not) in the decisions and their implementation. When we just let Americans make all the decisions, we end up with the sorry spectacle of American Facebook lawyers appealing in an American court to the superior wisdom of the US Congress in interpreting American free speech values in deciding whether or not it turns over evidence of a genocide in Myanmar to The Gambia to use in the International Court of Justice.

[26] See references at Kemp (1999, 967–68). For a particularly interesting discussion focusing on the negligence context and discussing other contrasting areas, see Gergen (1999).

[27] Cf. Twitter rules, enforcement philosophy, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy, visited July 27, 2021; https://perma.cc/LP4V-U5ZD.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

Moreover, work in fields like pragmatist epistemology can lead us to believe that colonial knowledge is also defective knowledge in terms of instrumental goals. For example, writing about moral philosophy – but with arguments equally applicable to corporate reasoning about things like content moderation – Elisabeth Anderson (2015) argues, drawing on classical pragmatists such as John Dewey, that reasoning processes under contexts of social hierarchy are tainted by the self-serving biases of the powerful, who have psychological dispositions to justify their own places in society. If this is right, and it certainly seems plausible, then it would seem that platform decision-making dominated by Americans and citizens of other wealthy liberal democracies would contain a built-in bias toward interpreting the needs of other societies in terms of the ideologies that they associate with their own success –for example, the alleged advantages of aggressive enforcement of intellectual property rights or less aggressive and facially neutral or "colorblind" enforcement of hate speech restrictions. This brings together the question of competence in terms of interpretation and the questions of morality in terms of inclusion: The same processes of inclusion that can make it possible for platforms to understand what a given utterance of maybe-hate speech means can also legitimate their judgment in determining what to do about it.

3.4 THE POLYCENTRIC MECHANICS OF DECENTRALIZED PARTICIPATION

The Myanmar case that leads this chapter focuses on enforcement – hate speech and incitement were already forbidden on Facebook, but lack of knowledge hampered enforcement efforts with tragic consequences. However, ordinary people on the peripheries also have useful knowledge about local needs that can bear on the making of the rules that are to be enforced. Human behavior generates endless novelty, including, alas, endless ways to cause harm to others or to shared resources; the effectiveness of policy solutions to shared problems can also depend on access to idiosyncratic knowledge as well as the epistemic benefits of a diverse community of reasoners (Cf. Hong and Page 2004). The high-speed evolution of behavior on the internet as well as the diversity of individual and social vulnerabilities suggests that the capacity of policy to adapt to novel behavior will require the ability to deploy knowledge far on the periphery from the people in San Francisco who currently write the rules for contemporary networked platforms.

Some scholars have, for that reason, advocated the aggressive decentralization of social media content moderation. For example, Wilson and Land (2021, 1074) argue that companies ought to create local teams in each country to write and enforce local rules.[28]

[28] To some extent, with the slow-boiling fall of Twitter, we may currently (as of this writing) be seeing experiments in far more radical decentralization, in which the federated Mastodon network creates (in effect) a market for moderation among interoperable microblogging platforms.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

I agree with Wilson and Land, but they perhaps do not go far enough. Company employees are still fundamentally implementing company norms and are still by some necessity socially distant from actual users and have at least partly conflicting interests. Moreover, discrete company employees are readily identifiable, and hence potentially vulnerable to government coercion, in particular in the countries where there are the most dangers of severe human rights violations of the sorts that Wilson and Land are concerned about, such as the Myanmar genocide.

One variation on the democratic strategy to mediate knowledge problems in the context of scale and diversity has, in the municipal and resource governance literature, gone by the name of "polycentricity." Associated with Elinor and Vincent Ostrom and the Bloomington school of political economy, polycentric organizations of government emphasize multiple levels operating at different scales with a degree of independence from one another as a method of generating policy innovation.[29] The notion of polycentricity seems to me to gel nicely with Ober's analysis of Athenian government, which emphasizes the capacity of democratic institutions that integrate diverse citizens with multiple sub-city affiliations and identities in multiple interactive institutional arrangements. In effect, the various kinds of possible governance actions of the Council, the Assembly, individual magistrates, and the citizen juries may have added this kind of polycentric character to the Athenian system. This, in turn, may have created a second level of incentives for knowledge aggregation, as citizens occupying these different institutional roles had to be brought into interaction with one another to reconcile their decisions, and citizens carried out different kinds of institutional roles calling for different kinds of knowledge and interactions when serving as a member of each of these groups.

Thus, it may behoove the designers of platform governance systems to consider what a polycentric system (or set of systems) of platform governance would look like. For example, could linguistic and subcultural communities on highly diverse platforms be given some regulatory autonomy, in order that they may both innovate in rules *and* transmit information to the center about developing idioms?

Several key insights from polycentrism theory are particularly relevant. The key proposition of the theory is that the grim predictions of the "tragedy of the commons" account of public resources, according to which users of a commons tend to defect from efforts to fairly allocate and limit overuse of that resource in the absence of external enforcement (or private property rights) (Hardin 1968), do not always empirically hold. Rather, in the real world, we see people successfully managing common resources all the time, if certain favorable conditions hold.

While platforms may or may not formally meet the definition of a common pool resource, they certainly share key features with them. In particular, they share the core incentive dilemma that users are often inclined to overuse and hence deplete

[29] See, generally, Aligica and Tarko (2012).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

shared resources. For example, a user who posts "clickbait" on a social media site effectively overuses a common pool of attention and credibility that is shared by other users who also desire engagement with their content.[30] Hence, it may be useful to draw on this research in understanding the likely structural conditions of successful platform knowledge and incentive management.[31]

One of the key favorable conditions is the existence of mutual trust in shared compliance with joint resource-management arrangements (Ostrom 2009, 11). As Aristotle recognized first, however, this kind of trust is more difficult in larger scale governance organizations.[32] At the same time, large-scale governance organizations do a better job of managing externalities in resource governance, because they are more capable of including all people who might be affected by both positive and negative externalities within the ambit both of collective methods of decision-making and of enforcement authority.

This tension creates obvious problems for local resources that have significant externalities, or where individual actions have both substantial and important local externalities and still significant but less substantial distant externalities. Consider environmental regulation: The most serious consequences as well as benefits to pollution may be within a single city, as the residents of that city capture most of the economic benefits from local industrial activity, have to breathe most of the particulates spewed out by it, and so forth – however, activity in the city may also contribute to pollution problems over a much broader area.

As the name suggests, the solution proposed by theorists who focus on polycentricity – which has been particularly influential in environmental governance – is, roughly speaking, "both/and." That is, governance organizations on the local level

[30] Benkler, Faris, and Roberts (2018), 288 analogize – accurately, I think – clickbait publishers on social media to polluters; cf. Morell (2014) who interprets at least one early social media platform – the photo sharing service Flickr – as a knowledge commons in the sense given by Frischmann, Madison, and Strandburg (2014). Kiesler et al. (2011), 129–130 helpfully articulate a number of Ostrom-derived design principles which even today platform operators ignore at their peril. For example, they recommend rate limiting, Kiesler et al. (2011), 151, 136–38, which companies many years later rediscovered through, for example, limits on WhatsApp forwarding to reduce virality; Hern (2020).

[31] In the existing scholarly literature, Frey, Krafft, and Keegan (2019) argue, based on Ostrom's work, for the role of participatory design in regulating platforms. They bring together the notion of a "constitutional layer" for "digital institutions" (a concept including, but broader than, platforms) with a "participatory design" tradition in human–computer interaction to argue for the advantages of low-level (i.e., ordinary person, or ordinary user) participation in the basic governing processes of such institutions (including rulemaking). The argument for these design decisions rests on ideas similar to those discussed in this book thus far – the authors observe, for example, that Colorado's cannabis regulation software failed to adapt to its environment because of the state's inability to receive information and feedback from the peripheries. Another key recent work is Forte, Larco, and Bruckman (2009), who identify the de facto devolution of certain kinds of rulemaking (such as stylistic norms) from central Wikipedia into "Wikiprojects" as an example of Ostrom-esque polycentricity.

[32] Aristotle suggested that the maximum size for a city was constrained by the capacity for all citizens to know one another's characters (Aristot. Pol. 7.1326b). Ostrom (2010a, 661) similarly points out that trust is easier to achieve in face-to-face contexts.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

as well as on higher levels are called for. Moreover, at the local level, theorists of polycentricity have identified that adjacent and even overlapping jurisdictions over people and activities may be beneficial rather than harmful, insofar as such arrangements provide for greater learning and competitive benefits: People under one jurisdiction can learn from the experiences of the next, and under some circumstances (such as in municipal service provision), jurisdictions can helpfully compete with one another to improve outcomes (Ostrom 2009, 33–34).[33]

Hiller and Shakelford (2018, 23–25) helpfully summarize the idea behind polycentric resource governance:

> Such a system is said to be polycentric if it has "many independent elements [that] are capable of mutual adjustment for ordering their relationships with one another within a general system of rules." Another definition of polycentric governance includes this concept's emphasis on "many decision centers having limited and autonomous prerogatives and operating under an overarching set of rules." A polycentric system, therefore, is not dependent on top-down government regulation, although it can include regulatory aspects. Instead, it is an organization of actors and method of governance that is multilayered yet interactive, independent yet networked, reinforcing and complementary. A polycentric system is complex, with differing rules and norms depending on the domain, providing for interaction between layers and among participants and allowing for "mutual monitoring, learning, and adaptation of better strategies over time." [...] Professor Ostrom also emphasized organizational spontaneity, meaning "patterns of organization within a polycentric system will be self-generating or self-organizing," and "individuals acting at all levels will have the incentives to create or institute appropriate patterns of ordered relationships." Factors necessary for self-ordering include freedom to participate, rules, enforcement, and adaptation.[34]

A background paper on climate change which Elinor Ostrom (2009) wrote for a World Bank report is particularly instructive, as platform governance seems importantly similar to the way she describes climate change governance. In particular, in both contexts, unwanted conduct creates both local and global externalities, and hence can – at least in principle – be managed by both global and local actors.

Ostrom (2009, 16) notes that communities can act to mitigate air pollution in the aid of their own local air quality, and hence the incentives to address local externalities can also help mitigate global externalities.[35] Now consider that within local

[33] One digital example comes from the Forte et al. (2009, 63) paper, who note that when various wikiprojects have overlapping jurisdictions, it serves as an opportunity for discursive dispute resolution.

[34] Hiller and Shackelford (2018, 24–25), quoting, variously, Koontz et al. (2015); Aligica and Tarko (2012); Ostrom (2010b). The reader may notice that the story of interdependence and effective governance through bottom-up development in this summary sounds remarkably similar, at a high level, to Jacobs's understanding of a successful neighborhood.

[35] In the context of social media, of course, "local" is a bit of a shaky concept. Geographic localities are obviously relevant, but so are what, for lack of a better word, we might call "affinity localities" like #BlackTwitter – groups connected by a shared identity as well as a relative density of network

groups on social media, efforts to mitigate their own vulnerability to low-quality content might have beneficial spillover effects. For a concrete example, a number of the 2016 Russian attacks on US politics via social media were locally focused, such as an infamous protest in Houston instigated by the Internet Research Agency (Riedl et al. 2021). Such events cause local harms to the political groups thus manipulated as well as broader political harms; it may be possible for platforms to provide greater affordances permitting, for example, people invited to local events to verify the identities of other participants and the organizers in order to permit local policing of such inauthenticity.[36]

Similarly, Ostrom (2009, 23–27) notes that in the climate change context, large-scale efforts to control pollution have often had difficulties recognizing local variation. But a significant part of the lesson of this chapter is that large-scale efforts to control platform information pollution have also failed to handle local variation. Thus, we can borrow Ostrom's suggestion that it may be beneficial to promote both local regulation and global regulation, to obtain the benefits of interpersonal trust and communication as well as learning in local regulation and permit the knowledge generated by localities to propagate out to the global level through, for example, observations of what works on a local scale and local selection into effective regulatory arrangements.

However, the literature also suggests that such cross-scale interactions must be designed appropriately, they cannot be left to chance. For example, creating overarching governance arrangements to permit lower-level actors to interact also introduces the risk of suppressing the system's ability to accommodate local variation (Young 2002, 283). Currently, we might understand platform governance as characterized by harmful interactions between governance scales. Companies generally start with a Silicon Valley-libertarian approach to user behavior both on the social media side (in terms of "free speech") and on the transactional side (in terms of openness to all comers, for example, shoddy products purveyed by drop-shippers with dubious advertising claims); with ad hoc changes to rules as they discover behavior which either harms their interests or as they are subject to pressure from

interconnections or clustering which have for that reason similar properties to groups of people who share physical space offline. In this chapter, when I say "local" I mean both the geographic sense and the affinity sense unless there is some obvious reason to limit the discussion to one or the other.

[36] Thus, Riedl et al. (2021) report that nobody – participants, counterprotestors, or journalists – could figure out the identity of the leaders of the protest in Houston. While the Houston example is problematic because the IRA-organized event was a white supremacist protest in front of a Muslim center – and we presumably do not want to make it easier for white supremacists to organize and manage their internal affairs – the IRA also notoriously targeted much more sympathetic groups and ideologies, such as Black empowerment and liberation. See Senate Intelligence Committee Report on Russian Active Measures Campaigns and Interference in the 2012 U.S. Election, v. 2, www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf, 38–39. I give the Houston example particular attention simply because the Riedl et al. study gives us the most insight into the failures of local self-defense against this manipulation.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

governments. Governments attempt to control conduct affecting their territories in a variety of ways, such as by controlling political speech which they do not like or defending intellectual property rights against counterfeiters on transactional platforms, but due to the frequent extraterritorial source of the behavior they wish to regulate, they typically find themselves operating via pressuring platforms rather than regulating directly. And users have little to no direct self-regulatory capacity, with the exception of some surfaces within social platforms with built-in self-regulation affordances (such as Discord groups and subreddits); likewise there is no international infrastructure to permit the conflicting claims of states to be effectively reconciled, or to provide for any formal engagement with civil society, indigenous peoples, and other organized nonstate actors outside of legislative processes.

At a minimum, experiments in what some commons scholars have called "co-management," in which higher-level and lower-level entities are jointly responsible for governing a resource (e.g., Berkes 2002), might involve, for example, some degree of devolution to groups of actual users of sufficient power over platform regulations to permit them to have some meaningful negotiating power with platform companies and governments. A part of the purpose of Chapter 6 is to suggest one way of implementing such an idea.

A closely related approach appears in a deeply insightful recent paper by political theorist Jennifer Forestal (2021a, on similar lines, see also 2021b, 2017). In it, she gives what amounts to a Deweyian version of the polycentrism theory, but focused less on governance and more on the structure of discourse on social media. Forestal argues, based on the theory of propaganda, that vulnerabilities to disinformation inhere in the underlying social structure of a population; in particular, its failure to be organized into salient but overlapping social groups which permit people to *both* engage in discussion on terms of shared interest and critically interact across those interests. While she does not use the term, the concept of "social capital" seems to me to serve as a good shorthand for such a propaganda-resistant social structure. Forestal compares Facebook's News Feed plus groups structure to Reddit's subreddit structure, ultimately arguing that the latter appears to be more successfully at promoting this sort of social capital.

In this context, one important lesson from bringing Forestal and Ostrom together is a recurrent point from the discussion in Chapter 1 of the way that the phenomenon of virality reveals a kind of false division between governance and product design. For the same overlapping but discrete groupings which might permit users to self-govern the content which they are producing and consuming can also, on Forestal's entirely plausible argument, improve their resilience against it in their capacities as mere users as well as democratic citizens. In view of Ostrom's insight, noted above, that polycentric governance tends to be "self-generating," we might hypothesize that a platform that is organized to facilitate the growth of social capital in terms of internal discursive health would also tend to promote the growth of groups suitable for a formal governance role. Hence, for example, existing subreddits might be recruited

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

to have more formal governance roles. An example of how such a development might come about is described in Chapter 6.

3.5 A DESIGN CRITERION FOR ANY POLYCENTRIC PLATFORM SYSTEM: ADAPTIVE CAPACITY

In addition to *distance* (in a geographic, cultural, or network sense), knowledge must also be managed over different *timescales*. Well-recognized strategies of governing in complex environments recognize that reliably predicting the outcomes of governance systems is often impossible in part because change is too swift, and that a core design criterion of any such system is to build the capacity for swift error correction in from the start.

An emerging area of social science and ecology that may characterize interactions on platforms – particularly, but not necessarily exclusively, the social media kind – is the notion of a complex adaptive system.[37] The "complexity" in such systems refers to the notion that assessing the entire system becomes difficult, including assessing the causal relationships between individual elements (or interventions on them) and the higher-level properties or other elements; actions can have "nonlinear" or "emergent" consequences (Miller and Page 2007, 3, 10, 27–28).[38] Miller and Page (2007, 233) observe that complexity tends to appear in systems that feature "heterogeneity, adaptation, local interactions, feedback, and externalities," which seem to me to generally characterize behavior on social media networks, and perhaps platforms as a whole. On such platforms, there are numerous actors, both institutional and individual (including ordinary users, companies, and even nation-states), who shape their behavior partly in response to the responses others give to their behavior, and so forth, in multi-directional feedback systems. One example is the perennial arms race between the "search engine optimization" industry and Google, as well as the ways in which content producers and consumers respond to changes in social media feed algorithms. Such feedback effects only increase when platforms interact with other social systems like markets or elections, as when, for example, r/WallStreetBets created massive shifts in the price of GameStop and other stocks, until controversial industry responses put a stop to it (Mezrich 2021). Heterogeneity and externalities are obviously present in such systems for the reasons discussed in the rest of this book.

This suggests that the choice of institutions to facilitate such governance must be particularly attentive to challenges involving understanding both local and global

[37] Cf. Lymperopoulos and Ioannou (2016), arguing as much.

[38] I understand "nonlinear" in this context to refer to a kind of disproportion rooted in the conjunction of feedback loops and extensive interconnections among agents, in which one seemingly small intervention can lead to chains of adaptive responses from the underlying agents, and hence surprising consequences, including in distant locations. "Emergent" describes "aggregate properties that are not directly tied to agent details" (Miller and Page 2007, 53).

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

effects of interventions (including seemingly small interventions), managing those effects across diverse sets of actors, and adapting to the discovery of surprising behavior – both positive and negative – in response to those interventions.

Within a polycentric governance framework, another potentially useful theoretical construct drawn from ecosystem scholarship that particularly focuses on complex systems is the notion of adaptive management. In a recent article on the management of the boundary between urban and wilderness areas, Craig and Ruhl (2020) helpfully describe both the contexts in which such strategies are likely to be most useful and their basic features. On their account, adaptive management is a "formal, structured decision process [which] involves a 'setup' phase, during which the decision-making actor specifies stakeholder involvement, management objectives, management actions, models, and monitoring plans, followed by an 'iterative' phase, during which the actor specifies the decision-making process, follow-up monitoring, assessment, and feedback" (Craig and Ruhl 2020, 616). That system is "resource-intensive" in view of the continuing infrastructure for monitoring and decision-making that it requires, especially in light of the fact that "external stakeholder engagement" is necessary to aggregate the kinds of information necessary for adaptation, but at the same time must be carefully crafted to avoid the risk of "suffocat[ing] the iterative decision process" by imposing excessive impediments to change (Craig and Ruhl 2020, 616–17).

Adaptive management is appropriate for systems that have several properties, including "chang[ing] dynamically over time," high uncertainty about the harms which may be caused by interventions on them, high "controllability," that is, there are lots of manipulations we can actually engage in, and a low risk of "irreversible transformation" (Craig and Ruhl 2020, 616). Under such circumstances, learning-oriented management strategies are called for in view of the fact that such strategies permit quick iteration (accommodating the changing nature of a system) in a context where such iteration is possible (because of high controllability) and minimally harmful (because changes are not irreversible); such iteration ameliorates the characteristic problem of such systems, namely the difficulty of predicting consequences, essentially by experimenting and seeing what happens. One characteristic type of system which is likely to meet the criteria described above is a complex adaptive system, one in which interdependent elements with feedback effects among themselves lead to emergent patterns of behavior and structure (Craig and Ruhl 2020, 614).

Because platforms probably constitute complex adaptive systems, and because a lack of information as to what is going on at the fringes of a platform ecosystem – and hence, by implication, as to the consequences of platform design choices and interventions on governance processes and rules – is a core challenge in platform governance, adaptive management may offer a particularly plausible toolkit to thinking about platform governance.

One challenge for the notion of adaptive management is that catastrophic consequences to experimental changes may indeed be possible – certainly existing design features of social media platforms have led to or contributed to catastrophic

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

consequences, such as the genocide in Myanmar and electoral distortions in the United States, Britain, and the Philippines. However, this may not entail the abandonment of adaptive ideas, but rather suggest that any form of learning-oriented governance must come with guardrails, such as a bias toward discovery of the sorts of information likely to point to catastrophic consequences in time to intervene ahead of those consequences. Moreover, given that our existing social media platforms are *already* leading to catastrophic consequences, if it is true that predicting the effects of interventions is impossible (because of complexity), then it may be that the probability of catastrophe would be reduced under a higher rate of experimental interventions (or it may not be; it cannot be known).[39]

Another consideration at play in deliberating on the adoption of adaptive management as an explicit platform governance strategy is the role of learning capacity within the underlying system. As Doremus (2011, 1471–72) argues, not all enterprises of governance offer a meaningful prospect for learning; in particular, if feedback on governance experiments is insufficiently fast, learning may occur insufficiently quickly to permit knowledge to be acted on. In the platform context, a related worry may arise, namely that due to the scale and diversity characteristic of platforms, the environment in which governance decisions are undertaken may change so quickly that, even if learning is possible, learning isn't fast enough to permit adaptation before it is, in effect, mooted by new environmental changes (cf. Doremus 2011, 1474 on the possibility of confounding changes in the environment). This suggests that any adaptive governance interventions must be accompanied by close attention to the speed at which knowledge is moved to the site at which adaptations are required. In the system described in Chapter 6 of this volume, this may suggest a bias toward giving local governance groups or councils more direct control over local platform outcomes, in order to permit swifter incorporation of new information close to the site of its impact.

In complex systems, adaptive governance arrangements may emerge rather than being imposed in the top-down fashion described by Craig and Ruhl; one way to represent the task of a governor in such a system is to create the conditions under which such adaptive institutions may emerge in ways that facilitate the achievement of public goals (Cosens et al. 2021). As Cosens et al. recognize, there is an overlap between the ideas of adaptive governance, new governance (which, as described in the introduction to this volume, focuses on multistakeholder inclusion), and Bloomington School polycentric solutions.[40] In their words:

[39] One plausible way to model the problem is as a landscape of possible platform design and governance configurations, some proportion of them which pose catastrophic risk; it may be that we are currently in a region of that landscape in which many configurations lead to such risks, but that other regions lead to lower risks. But this depends on one's priors on the distribution of catastrophic risk, and there does not seem to me to be much to recommend any particular view.

[40] I read the shift in emphasis over some of these papers from "adaptive management" to "adaptive governance" as perhaps partly reflecting a greater bottom-up character – however, I might just be over-reading a minor linguistic difference.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press

> New governance and adaptive governance both include as essential features bottom-up self-organization and collaboration; public, private, and public–private networks; and multiple nested centers of authority (i.e., polycentricity). Adaptive governance literature calls out the presence of processes to manage uncertainty and mechanisms for learning and incremental adaptation. Both emergent forms of governance respond to increased connectivity and complexity and the corresponding need for contextualization and adaptation. They appear able to navigate and maintain overall social stability when system trajectory is uncertain and fraught with surprise. In short, they represent societal responses to complexity.[41]

In many ways, this version of adaptive governance which offers specific attention to the notion of preserving the public character of multistakeholder models while consciously adapting to social complexity, sounds remarkably like a reprise of John Dewey.[42] Elizabeth Anderson (2006, 13–14) describes Dewey's conception of democracy as analogous to the scientific method, in which public policies are framed in deliberation among diverse and inclusive groups and then tested to determine their outcomes, with democratic institutions like free speech and periodic elections providing information both for hypothesis generation and feedback/observation. What the idea of adaptive governance adds is a healthy skepticism of predictive methods of policy framing, and hence a bias toward governance systems that can generate and change ideas and experiments at a higher rate and are more responsive to feedback. In a complex system with actors across multiple locations, the vigorous use of democratic polycentricity may, I hypothesize, facilitate learning in the same way that federalism understood as the creation of political laboratories does, that is, by creating additional sources of change and sites of observation (with measures to swiftly disperse those observations across the governing network) while permitting simultaneous experimentation with novel options.

Chapter 6 of this book will sketch out a system of bottom-up platform governance designed with the capacity to adapt to change as a first-order goal. But before we get there, we must first look in the opposite direction: Governance requires not only innovation but also constraint, and many of the problems generated by platforms are arguably attributable to the lack of constraint, particularly of company executives. Chapter 4 considers the ways in which political states facilitate that constraint, so that we may later design institutions that can meet both needs: Innovation from below and constraint at the top.

[41] Cosens et al. (2021, 4), internal citations omitted.

[42] Doremus (2011, 1464 n. 35) notes that a number of prior scholars have read adaptive management in the Deweyian tradition.

https://doi.org/10.1017/9781108975438.004 Published online by Cambridge University Press