A HOOVER INSTITUTION ESSAY

Verified Accountability

SELF-REGULATION OF CONTENT MODERATION AS AN ANSWER TO THE SPECIAL PROBLEMS OF SPEECH REGULATION

EVELYN DOUEK

Aegis Series Paper No. 1903

The "techlash" of the past few years represents a moment of quasi-constitutional upheaval for the internet. The way a few private companies have been "governing" large parts of the digital world has suffered a crisis of legitimacy. Calls to find mechanisms to limit the arbitrary exercise of power online have gained new urgency. This task of "digital constitutionalism" is one of the great projects of the coming decades.[1] It is especially pressing in the context of content moderation—platforms' practice of designing and enforcing rules for what they allow to be posted on their services.[2] Historical forms of public and private governance offer limited guidance. Platforms are not nation-states.[3] But the unprecedented power that major tech platforms wield over individual rights and public discourse also differentiates them from corporations as we have known them. This outsized impact on rights so central to democracy has led to demands for greater accountability for decisions about online speech, but extensive government intervention can be a poor tool for achieving this goal. Government involvement in speech regulation is uniquely pernicious, and the cure for the problems with content moderation should not be worse than the disease.

Instead, platforms need to innovate to create new forms of self-regulation to meet the demands of the moment. To date, the most developed proposal to create an independent constraint on its content moderation is that of Facebook's Oversight Board, a court-like body that will hear appeals about the most difficult and important content moderation decisions Facebook makes and give public reasons for its decisions.[4] This essay takes the Oversight Board concept as the basis of a possible model in examining the promises and limitations of platform self-regulation. While this paper's focus is on the benefits of an independent appeals body, to be truly effective any such body needs to be embedded in an entire system of governance. As Facebook has itself noted, this is a project that will take years.[5] But the task is also urgent given its importance and real-world effects.[6]

Semi-independent and transparent self-regulatory oversight mechanisms offer significant advantages, not only over the current delegitimized governance structures but also in absolute terms. As the actors closest to the front line, platforms will always need to play a significant role in drawing lines for online speech, given the high-volume, fast-moving, and context-dependent nature of the decisions involved.[7] A recent French government report acknowledged the benefits of this responsiveness and flexibility in endorsing a model of government regulation that "capitaliz[es] on this self-regulatory approach already being used by the platforms, by expanding and legitimising it."[8] This expansion and legitimacy

can come from internal oversight, which can create a forum for the public contestation of platform rules and their implementation. But it is also true that self-regulatory solutions are likely to be a significant disappointment to many. They will not be able to meet the current expansive demands for due process and transparency in most content moderation decisions. Nor will they be able to create global norms about the appropriate limits of freedom of expression. But these goals set unrealistic benchmarks.

This essay looks first at the puzzle of why a private company might engage in a kind of constitutionalism and the demands that such projects are designed to meet. I then turn to the limitations of these measures before explaining their benefits and why they will be an important part of the future of online governance.

Old Governance and New Speech Realities

This moment demands innovation in platform governance for two main reasons: the growing momentum behind the belief that private actors can and should be called on to provide greater public accountability for their content moderation practices, and the peculiar needs of legitimate speech governance that necessarily shape the form that such accountability must take. Together, this creates a need for a new form of governance that can legitimize the unprecedented power that private tech platforms have over public discourse.

Private Public Accountability

Charges that platforms' content moderation is "illegitimate" are diverse and hard to distill into a simple objection.[9] One difficulty is that notions of public accountability and constitutionalism initially seem out of place in the context of a private company serving willing users. But the major tech companies create speech rules, enforce them, curate and censor content, and resolve user disputes, all in the course of moderating some of the most important channels of information in the world. A major aspect of the techlash has been growing awareness that tech companies lack accountability as they exercise this substantial power.[10] Their decisions have come to be viewed as inconsistent and arbitrary, and therefore illegitimate.[11] It is this demand for public accountability that creates the need for platform self-regulatory innovation, even if it is not currently legally required.

Indeed, platforms have already been systematizing their speech rules and slowly providing carefully circumscribed transparency in an effort to garner greater public legitimacy. This is not pure public-mindedness. Facebook's head of policy management has written: "Simply put, there are business reasons that a big social media company must pay attention to what the world thinks of its speech rules."[12] A Facebook-commissioned report by a group of independent academics called the Data Transparency Advisory Group (DTAG) explained:

> Facebook has considerable capacity to manage its content regulation process in a top down manner which pays minimal attention to users' views. However, the very existence of the [Community Standards Enforcement Report] highlights the recognition that public views

> about Facebook and its attitude about the public matter. They matter for the individual user both because disgruntled users find ways to circumvent rules, for example opening multiple accounts. In addition, unhappy customers are less likely to use the site and more likely to seek alternatives to it.[13]

But it is one thing to say that, descriptively, platforms have been developing a kind of platform law that resembles the common law,[14] or that Facebook's Oversight Board is a kind of constitutionalism.[15] It is another thing altogether to suggest that it is actually required or normatively desirable to redesign platform governance to create constraints on the way platforms exercise their power. This latter idea can seem incoherent to those acculturated in American legal norms.

In the United States, the state action doctrine means that constitutional rights only apply to actions taken under the auspices of government.[16] American courts have been reluctant to expand the reach of state action beyond narrow bounds,[17] even as the doctrine is increasingly strained by more privatization and government outsourcing, use of arbitration clauses by companies, and reliance on the private infrastructure that makes much public online life possible.[18] The Supreme Court in a June 2019 decision made clear that it has little appetite for applying constitutional constraints to tech platforms when it emphasized that "merely hosting speech by others is not a traditional, exclusive public function and does not alone transform public entities into state actors subject to First Amendment constraints."[19] This attitude reflects more than mere constitutional doctrine; it is an underlying ethos that focuses on government power as the primary threat to individual liberty. This same ethos animated John Perry Barlow's famous Declaration of the Independence of Cyberspace, directed to the "Governments of the Industrial World."[20] And it reflects a core element of the ideal of liberal autonomy that private actors' reasons for acting should be free from public scrutiny.[21] From this perspective, a private company's actions do not necessarily need public legitimacy beyond what is commercially desirable, and there is no reason to assume that the relationship between legitimacy and profitability is linear.[22]

But this is not the only way to understand rights. Many other countries give "horizontal effect" to rights, reflecting greater recognition of the threats that nongovernmental actors pose to rights and a stronger commitment to social democratic norms.[23] The United Nations Guiding Principles on Business and Human Rights, a nonbinding but widely endorsed statement of international law, also recognizes the obligation of business enterprises to avoid infringing on the human rights of others.[24] This "horizontal" model is the intuition behind the (mistaken) complaints by users that platforms are infringing their First Amendment rights when they moderate. It is also the intuition behind the claims that tech firms possess a kind of "sovereignty"[25] or are "public utilities,"[26] the equivalent of the modern public square,[27] or an old company town.[28] None of these analogies fit easily, and all are attempts to shoehorn new problems into more familiar solutions,[29] with potentially detrimental consequences.[30] But they do underline the problem that self-regulatory innovations aim to meet: a growing sense that governments are no longer the only, or even the primary, threat to rights and that when

private actors construct ecosystems to manage public rights, these ecosystems should bear the characteristics of public accountability, due process, and the rule of law.

This cultural difference in the understanding of rights means global platforms will face a large number of jurisdictions where governments and users are more accustomed to requiring private actors to conform to public accountability and rule of law norms.[31] Shifting public sentiment, congressional hearings, and even legislative proposals suggest that the tech giants will increasingly face this pressure in the United States too. Of course, there continue to be practical obstacles to holding corporations accountable for rights infringements in many cases.[32] But the trend of the global discourse about tech governance is clear. Platforms have an interest in being proactive about the shape of that conversation by creating new forms of self-regulation that meet the growing demands for better content moderation performance and accountability.

The need for better systems is exacerbated by laws that increasingly impose legal responsibility on platforms for their moderation of public discourse. For example, Germany's Network Enforcement Act (NetzDG) delegates the interpretation and enforcement of speech regulations to platforms by imposing large fines on platforms that fail to take down certain types of illegal content within short time frames.[33] The right to be forgotten in Europe represents another example that requires search engines to make judgments about what information is in the public interest.[34] These kinds of laws (of which more seem to be on the horizon) create greater legal responsibility for company content moderation on controversial issues. The concurrent and increasing concern about the legitimacy of these decisions by platforms creates the need for a more mature form of content moderation that factors accountability and transparency into its design.

It is worth noting that at present there is nothing legally entrenching platforms' proposed self-regulatory reforms. They currently remain voluntary, and platforms may decide to push back against the growing demands for public accountability in their content moderation ecosystems, reasserting their rights to manage their commercial operations as they please.[35] But this may not remain the case. As noted above, France has indicated interest in a regulatory model in which increased self-regulation will be mandated for platforms that meet certain criteria. But voluntary initiatives by platforms today can shape the form of these future mandates. The members of the French mission who endorsed this model did so based on "the progress made in the last 12 months by an operator such as Facebook,"[36] showing that self-imposed platform reform and regulatory reform can occur in dialogue with one another.[37] Whether this is viewed as regulatory capture or a virtuous cycle depends on ensuring that the self-regulation that results is effective.

Speech Is Special (and Especially Difficult)

There are a number of features of freedom of speech that make designing a legitimate system of speech governance especially difficult. Not every feature is unique to speech

rights, but combined these considerations show the ways that speech governance raises unique challenges.

First, as Frederick Schauer says, freedom of speech is "a somewhat different type of 'right'" because speech is positively advantageous and should be encouraged, not merely shielded from state intrusion.[38] This means that regulation of online platforms is different from historical forms of gatekeeper regulation, which typically focused on using gatekeeper liability to prevent underlying *misconduct* rather than the *promotion* of rights.[39]

The positive benefits of free speech also led to unique doctrinal developments: broader standing rights to allow vagueness and overbreadth challenges to statutes and concerns about "chilling effects."[40] These rules reflect the notion that, more than is the case for other rights, any restrictions on speech need to be clearly defined in advance. But this is made more difficult because the contours of freedom of speech are essentially contested. The frontiers of free speech are especially uneven,[41] differing from country to country.[42] Furthermore, what constitutes freedom of expression fundamentally differs in different contexts.[43] The fact that context matters so much in translating speech rules into concrete decisions in individual cases poses a unique challenge for regulation of speech on global platforms, which facilitate large volumes of speech across vastly different settings. As Facebook's head of policy management, Monika Bickert, described the problem:

> The practicality of implementing standards in communities this large simply requires a heavy hand from the companies. Even if online speech standards were set by an outside authority, the level of attention required to implement any standard at such a large scale means that companies must play a primary role in the ultimate decision to remove or leave on site any given piece of content.[44]

Free speech cases also involve the collision of different rights more often than other cases. Regulating expression, especially hate speech and sexually explicit materials, for example, creates conflicts between liberty and equality. One person's liberty of free expression has direct impacts on the dignitary interests of others. These clashes are "common and readily expressible as 'zero sum' situations."[45] These kinds of zero-sum controversies are especially contestable and controversial. Moreover, because they involve balancing different interests, they are ill-suited to unilateral, unaccountable, and unexplained decision making. Therefore, although freedom of speech is never absolute and can be restricted to take account of other sufficiently important interests, these restrictions should be publicly explained and provided for in a way that limits discretion.[46]

But private actors are not well-equipped to evaluate the interests that law typically permits freedom of expression to be limited in the name of, such as national security or public safety.[47] By contrast, platforms have a legitimate interest in regulating speech in ways that would never be permissible for a government. As Tarleton Gillespie put it, content

moderation is *the* commodity that platforms offer.[48] Making choices about speech is their business. The way they rank and present content is what attracts or repels users. A website focused on knitting should be free to decide it does not want to host certain political content that it determines does not align with its mission.[49] This difference causes problems in trying to apply existing rules and case law about speech, centered around governments, to private platforms that have a wide variety of business models and capacities.

Perhaps most relevantly (and problematically) for questions of internet governance, state involvement in any aspect of speech regulation is especially fraught. First Amendment doctrine in particular is deeply shaped by a distrust of government motives in speech regulation.[50] But other legal traditions have similar concerns. The French report, for example, noted the "special precautions" that need to be taken when public authorities become involved with speech regulation.[51] Particularly when it comes to regulating political speech, there is a strong suspicion that governments might act on illegitimate motives related to preservation of their own power.[52] Therefore, state regulation that aims to rein in private power over individual expression may itself undermine the democratic purposes of free speech.[53]

Finally, speech rights are also special because free speech is fundamentally a public-facing right. The rights of the individuals involved remain important, but speech is prized for the wider benefits it brings to society, and it is sometimes feared for its potential to create wider harms. The ramifications of any particular rule or decision for wider public discourse loom larger than they do in many other contests over rights.

To summarize: there is no "correct" understanding of free speech. Restrictions on free speech should be clear in advance, but they also need to be applied with special attention to context. Who sets speech rules is especially important, and the role of government needs to be carefully circumscribed. But outsourcing decisions so central to public discourse to unaccountable private actors also causes concern. These are the special challenges of speech governance.

As a result, regulatory frameworks imported from other contexts cause problems for speech governance. For example, multistakeholder initiatives about supply-chain integrity may benefit from opaque negotiations if assurances about the protection of commercial information make companies more willing to cooperate. But a lack of transparency in negotiating respect for freedom of speech is *intrinsically* inadequate.[54] Another example is the increasingly diverse ranges of online dispute resolution (ODR) systems that platforms use to resolve other conflicts, such as disputes between a buyer and seller in an online marketplace over whether a product sold matched its description. These ODR systems typically involve resolving disputes between two private parties on a platform,[55] instead of the more "public law" style of disputes that are involved in contesting the ruling of the platform "government." While many other ODR systems are increasingly automating their processes and removing human handlers from the dispute-management process,[56] the nature of speech disputes distinguishes them and makes this automated approach

intrinsically flawed.[57] Automated processes struggle with highly contextual decisions, and even when algorithms get the decisions "right," they cannot communicate them in a way that will be acceptable to the affected parties.

Understanding the peculiar problems of speech governance in this way provides insight into the role that an internal oversight body, such as Facebook's Oversight Board, could fulfill. It also illuminates the most effective role for government regulation. Governments cannot and should not micromanage rules for online speech. They can submit permissible bounds in a broad sense, in keeping with the restraints of their individual constitutions. But for three reasons, these rules will likely not provide an answer in most of the hard cases that platforms have to decide in liberal democracies.

First, platforms will impose restrictions on speech that go beyond what a government might be able to, whether for practical reasons (such as to prevent platforms becoming overwhelmed by spam or fake accounts) or because of their business model (such as platforms that wish to ban adult content, or communities that wish to provide a politics-free zone).

Second, government regulation could not manage the speed and scale at which platforms have to decide these questions. As Bickert notes, online speech regulation requires a heavy hand from platforms to translate rules into on-the-ground decisions in a dynamic environment. Governments may set broad standards, as in the cases of NetzDG or the right to be forgotten, but these only go so far. Online norms of discourse, memes, and coded language change by the day, or even the hour. Government processes are unlikely to be agile or responsive enough to manage the rapid evolution of online speech disputes, let alone apply them to millions of individual cases.

Third, even if constitutionally permissible or practically possible, the notion of governments having such extensive involvement in speech regulation is in deep tension with the underlying democratic purposes of free expression.

Therefore, government regulation of content moderation should focus on legitimizing the *processes* by which platforms make decisions about speech rather than targeting the *substance* of those decisions. Major platforms should be required to show that they are moderating in accordance with their publicly stated rules.[58] Platforms may be overtly political or politics free, but these values should be transparent to provide accountability. Ensuring that content moderation accords with public rules requires platforms to create systems that make content moderation more than the sum of individual ad hoc decisions. To ensure that platforms moderate speech systematically rather than arbitrarily, governments can impose disclosure obligations about the basis of platform decision making, as well as require internal oversight mechanisms that render platform decisions publicly accountable. These mechanisms can be audited and reviewed for effectiveness; public rules and an internal appeals body mean nothing without systems in place to carry out their mandates

consistently in other cases. We might call this a model of "verified accountability," where platforms have an obligation to make aspects of their governance transparent and accountable, while governments regulate to verify these commitments. Another benefit is it does not seek to impose one-size-fits-all obligations on platforms, which differ significantly. This is not merely a hypothetical idea for regulation: it is the idea behind the French report's recommendation of "expanding and legitimising" self-regulation to create greater trust and legitimacy for online rules.[59]

The question then becomes what "public accountability" means in the context of a private platform making speech decisions at scale. While internal oversight can improve the accountability of the currently mostly opaque systems of content moderation, it is important to be up front about the fact that there are a number of demands that these mechanisms will not be able to meet.

The Limits of Self-Regulatory Oversight

Defining success for self-regulatory content moderation requires realistic expectations and an understanding that there are many complaints about platforms' content moderation ecosystems that an individual oversight body or a self-regulatory system in general cannot address.

"Perfect" Process in Every Case

One of the most pervasive criticisms of the current way platforms moderate content is that their systems do not provide adequate due process to affected users.[60] Many legal systems provide an avenue for appeal in order to afford due process because appeals give complainants an opportunity to voice their grievances, have hearings, and receive some forms of explanation for their treatment.[61] But the scale of the largest platforms makes this expansive conception of due process impracticable. As Alex Stamos has noted, Facebook makes more content-moderation decisions in one day than the US justice system makes all year.[62] There are also difficulties in translating notions of "due process" that center around the exercise of *state* power to private actors. On the one hand, for example, requiring platforms to provide additional process does not give rise to the same questions of trade-offs in the use of public resources. On the other hand, the use of coercive power by the state to infringe on private rights raises very different questions than the exercise of power between two private parties. These differences are rarely acknowledged in calls for greater due process in content moderation.

Nevertheless, there are substantial benefits—including legitimacy—to providing a kind of procedural fairness (something akin to what we think of as "due process") online.[63] Scale alone does not justify jettisoning these benefits, and platforms are increasingly offering greater procedural protections to users. But *due* process does not mean *perfect* process. Conversations about due process in content moderation should more explicitly account for the fact that the right itself varies according to context. This more nuanced

discussion around what the "due" in "due process" means in the context of the scale of online platforms requires grappling with what kinds of errors society prefers.

Traditionally, US law has erected strong due process protections for speech rights, as against governments, requiring procedures used to show "the necessary sensitivity to freedom of expression" and acknowledgment of chilling concerns.[64] But the general principle remains that due process is flexible and depends on the circumstances.[65] Determining what process is due in any case requires consideration of the private interest affected (and, in the context of private platforms, the *multiple* private interests affected—both the platform's and the user's), the probable value of any additional procedural safeguards, and the burden that such additional safeguards would entail.[66] Critically, "procedural rules must always be designed as a system, in light of the overall goal of the programs."[67] Fundamentally, "the very nature of the due process inquiry indicates that the fundamental fairness of a particular procedure does not turn on the result obtained in any individual case; rather, 'procedural due process rules are shaped by the risk of error inherent in the truth-finding process as applied to the generality of cases.'"[68]

Discussion of process in the context of content moderation typically does not take this systemic view. The Santa Clara Principles developed by civil society organizations, for example, call for a meaningful opportunity for timely appeal—which includes the opportunity to present additional information and the provision of reasons for decision—for "any content removal or account suspension."[69] But Facebook removed 2.2 billion fake accounts in the first quarter of 2019 alone, as well as 1.8 billion pieces of content for spam violations.[70] People appealed 20.8 million of the decisions made in relation to spam. This is an average of around 231,000 appeals a day. Facebook takes action against content for violating its rules against spam more often than in any other category, but it is still only one category in its community standards. Overall, Facebook received nearly 25 million requests for appeal of content in the first quarter of 2019. The company does not release information on how many additional pieces of content it algorithmically reduced distribution of without taking down, an enforcement tool the company is increasingly using to deal with content that approaches the line of what it prohibits.[71] Focusing on suspensions should not obscure the fact that such down-ranking could have a very similar effect in practice on how many people see a post. These numbers make clear that requiring reasons in every case would overload even a vastly expanded content moderation workforce.[72]

It is not clear that a system that allowed an appeal with an opportunity to present additional information in all of these cases would serve the goals of greater due process in the system overall: there would be trade-offs in consistency between decisions, the timeliness of review, and the level of reasoning that could be provided, as well as in diverting resources to categories such as spam that could be devoted to other case categories, including hate speech and bullying. This is not unique to content moderation; in the design of all systems of procedural justice, there can be trade-offs between accuracy in

any individual case and overall systemic accuracy.[73] It may be that spam is a particular category of case where affording fewer procedural rights can be justified—after all, there is even government regulation against spam.[74] But this just illustrates the point that what "due process" means needs to be determined contextually. It is not enough to say that because speech rights are involved, robust due process protections should always be provided, particularly when private actors, and not the government, are involved. Of course, if different categories of content are treated differently, it will become more important to ensure that platforms are accurately categorizing their decisions. This is where self-regulatory auditing, enforced and verified by legal mechanisms, can play a key role.[75]

Explicit discussion has barely begun about how to draw lines between different ways platforms treat content, what kind of process should attend each case, and how procedural design affects the operation of the content moderation ecosystem as a whole. The UN special rapporteur for freedom of expression, David Kaye, for example, has acknowledged the need to find "scalable solutions" to the problem of content moderation, including establishing criteria for complaints that qualify for additional appeals, given how time-consuming and costly it would be to allow appeals on every content action.[76] A more sophisticated conversation about due process might start by more explicitly differentiating between the different categories of speech that content moderation implicates, accounting for the difficulty of making a correct decision in each category, the public importance of the underlying speech, and how people experience different kinds of decisions. The stark differences in how people feel about content moderation decisions in different categories can be seen in the differences of appeal rates. For example, when Facebook finds a piece of content to be hate speech or bullying, users appeal those decisions far more often than in other categories.[77] This could reflect the more difficult nature of the initial decision, but it also might reflect the greater grievance a user feels when they are said to have violated hate speech or bullying policies. Both possible reasons should factor in to the design of content moderation systems.

Another key factor that system design must account for is speed of decision making. Because content can go viral online within minutes, reducing the impact of harmful speech often requires a platform to remove it quickly. But this inevitably increases the error rate. A systematic understanding of due process might recognize that such temporary errors are an inevitable and acceptable trade-off against the need to impose rules within a time frame that makes them meaningful. Perhaps this means more willingness to accept occasional mistakes when educational or journalistic depictions of extremism are temporarily swept up in bans on extremist content, for example.[78] Or more consciously deciding that accidental censorship of more videos of first-person shooter video games may mean that first-person depictions of real-world violent attacks will be more reliably detected.[79] Currently, it is difficult to make overt decisions to accept a certain level of error. Instead, as Gillespie says, currently "one failure can incur enough public outrage to overshadow a million quiet successes."[80] But this results in the type of error that systems favor being chosen less consciously. Oversight can provide a forum for error explanation and more deliberate

choices between trade-offs involved in any system design. However, accepting errors also requires trust that companies will still work to reduce them and have systems in place to fix them. Again, this is where regulation and auditing can play a role.

In a dramatically changed speech environment where speech is cheap and attention is scarce,[81] and when the biggest threat to free speech may not be censorship but exploiting and distorting online discourse through speech itself,[82] the calculus of due process requires different conclusions to prior free speech case law that emphasized that the specialness of speech required more robust process. Requiring hearings for appeals on all content decisions might incentivize laxer substantive moderation rules, which could cause other unintended harms. Striking the correct balance between substantive justice, speed, scale, and accuracy requires a more conscious calibration of the level of process that is afforded to affected users. For this reason, self-regulatory oversight will not, and should not, be an answer to calls for expansive due process rights analogous to traditional court hearings about censorship.

Transparency in Every Case

Calls for greater transparency have become common in criticisms of content moderation, perhaps because transparency is often lauded in policy discourse as low-cost, simple, and easy. What is more, company objections can sound defensive and purely self-interested, as if there is something to hide. As discussed above, a measure of transparency is essential to creating accountability. Certain aspects of content moderation, such as the rules and values underlying platform decisions, need to be made public if platforms are to be held to them. The problem is that most calls for transparency are generalized and lack specific content.[83] But "transparency" can mean many things, not all of them equally beneficial, and abstract assumptions about its virtue do not advance the conversation about what exactly companies should be doing.[84] A simple example is the legitimate need to avoid adverse privacy or safety implications of providing too much transparency in cases of hate speech, bullying, or reports on dangerous organizations. Republishing and providing greater visibility to the content of hate speech or bullying messages as part of an effort at "transparency" may in effect do the work of the person the platform has decided should not be heard. Similarly, there is the need to avoid the "Streisand effect" in removing "doxing" posts or in right to be forgotten cases, where respect for an individual's privacy is the point. Because transparency can be an instrumental good for achieving greater accountability, the benefits and costs of particular transparency measures should be considered in a more specific sense, rather than by reference to abstract concepts such as "legitimacy" alone.[85] What is important is, again, a systemic focus that reveals more useful information than a focus on public transparency in each individual case.[86]

While there have been moves toward more transparency in recent years, the enforcement of platform rules currently operates largely in the shadows. The major platforms release aggregate numbers about enforcement actions taken against various types of content, but without examples or further information it is difficult to assess what these reports mean.

For example, when Facebook says it took down four million pieces of hate speech in the first quarter of 2019, more than in any previous quarter, and that this reflects "improvements and expansion" of their proactive detection methods, it is impossible to assess these claims.[87] There is no independent verification of numbers or whether what is being taken down actually meets the definition of "hate speech" in Facebook's Community Standards. As the French report puts it, "The persistent dissatisfaction of the public authorities can be explained in particular by their inability to assess the measurable reality and value of the self-regulation carried out by these operators, due to a lack of information validated by a trusted third party."[88] Without verification or intelligibility, aggregate reports become a form of transparency theater, deployed to ward off calls for greater accountability.[89] Validated numbers alone are not enough. Individual examples and sampling of enforcement decisions are necessary to illustrate what abstract platform rules mean in practice and the level of accuracy of enforcement decisions.

An independent oversight body can make abstract rules more comprehensible by providing examples in a small set of cases. By giving reasons in especially difficult cases, those that sit on the borderline of categories, the shape of the category as a whole becomes more understandable.

But when it comes to development of platform "case law,"[90] a single body cannot provide transparency in any but the smallest proportion of decisions given the sheer scale of platform operations. This is not necessarily a bad thing: privacy and safety concerns will often mean individual cases should be kept from public view. In right to be forgotten cases, disclosure of identifying information would violate Europe's General Data Protection Regulation.[91] But there then needs to be a mechanism for ensuring that in those cases platform rules are being applied with an acceptable level of consistency and without bias. Internal self-regulation or independent auditing can play a quality-assurance role here. An example is Facebook's ongoing civil-rights audit: an independent team of civil-rights experts were able to conduct an investigation into the enforcement of Facebook's hate speech policies.[92] Auditors were able to observe reviewers at work and examine samples of incorrect enforcement decisions. This process resulted in concrete recommendations for how Facebook could improve its moderation processes.

Accepting something less than radical transparency requires trust that there are systems in place to ensure fairness in those cases that are not visible. If accountability cannot be completely public, it should be verified through system testing. Self-regulatory oversight can help remedy the lack of trust created by the history of opacity and obfuscation in content moderation to date, but it will not create sweeping transparency in all cases.

Global Norms

Platforms often prefer one set of globally applicable content moderation norms, and their terms of service or community standards tend to be phrased in these terms. The claim is that global rules are necessary for reasons of "efficiency and the completeness of the

borderless experience."[93] Platforms no doubt have costs reasons to want to avoid a different set of rules for every jurisdiction. But an internal oversight body, even one that prioritizes diversity among its members,[94] is unlikely to be able to resolve what are likely intractable differences as to appropriate speech norms between different communities. Even if such a substantive reconciliation of differences were theoretically possible or normatively desirable, it is unlikely that any private company would be able to establish a body with sufficient legitimacy or credibility to bring this about.[95]

Instead, internal oversight bodies should focus not on harmonizing rules between communities but on bringing diverse perspectives to bear on how platform rules need to be applied differently in different contexts. The history of platform-content moderation has partly been one of learning about the need to accommodate local concerns, and the realization that US First Amendment norms are not universally applicable or appropriate.[96] In hard free speech cases, especially in areas such as political speech or hate speech, context is all-important. This means that broad rules—even international human rights norms—will have different applications in different situations.[97] While not to understate the structural and social harms of racism, the risk of imminent danger created by a racial slur may be very different in a Western democracy than in the context of a society with ongoing widespread ethnic violence, for example.

Internal oversight and self-regulation will also be unable to resolve tensions between proliferating local laws requiring different rules in different jurisdictions. How and when platforms apply these laws is going to be one of the defining questions for the future of free speech online. Where government demands are inconsistent with stakeholder values, independent oversight might help those stakeholders pressure platforms to push back by bringing attention to these cases. David Kaye also suggests that if companies ground refusals to accede to government demands in the language of international human-rights law, it might increase the credibility and force of those platforms' pushback.[98] Internal oversight could facilitate this process if it were charged with explaining how government rules were inconsistent with more universal norms. But the role for an oversight body in protecting against government overreach is limited. Because government interference in speech is considered especially pernicious, this is one area that transparency should be provided as a matter of course and not merely in response to individually initiated cases that require affected users to bear the burden of raising the issue. And to the extent that governments use coercive power to mandate certain content regulation through law, an internal oversight body cannot mandate noncompliance.

What Internal Oversight Can Do

Despite the limitations described above, internal oversight bodies can be a useful innovation and can significantly improve current content moderation ecosystems.[99] The exact design of such an institution of course matters a great deal. Facebook's Oversight Board is to date the only such proposed initiative, and the publicly available details about

how it will operate are still very high level.[100] Furthermore, given the diversity in platforms' services and structure, oversight bodies should be tailored to each platform's specific context. The discussion that follows is therefore based on an ideal type: an oversight body staffed by persons independent of the business arm of the platform that hears appeals about the platform's enforcement of its content moderation rules and gives public decisions affirming or modifying platform practice that are then implemented throughout the platform's content moderation ecosystem. This description obscures a lot of complexity: who can bring complaints and how they are chosen from the millions of potential cases are key among the details that greatly impact the effectiveness of the institution. Nevertheless, this description provides a useful starting point for identifying the advantages that such a body can bring, which then can guide specific design decisions.

These bodies can provide two main benefits: a "judicial"-style check on platform decisions can improve platform policies, and public reasoning can create greater acceptance of rules in a pluralistic community where disagreement about the substance of those rules is inevitable. An oversight body can also provide greater visibility to community norms, diffuse public pressure, and explain moderation decisions about government speech.

Improve Platform-Policy Processes

A "judicial"-style check on the "legislative" activities (i.e., the policy formation) of a company may improve the overall functioning of its system. Even an oversight body that can be overruled can help highlight blind spots and counteract inertia in the formulation of content rules.[101] Blind spots are created both by the fact that policy is often written under time constraints and with limited foresight about the full range of circumstances in which rules will be applied.[102] Inertia arises because changing rules is time consuming; in addition, when there is persistent disagreement in difficult cases, the status quo often prevails.[103] Judicial-style review counters these problems by creating opportunities to raise issues or address complications that arise when policies are applied to specific cases and by providing a way to disrupt the status quo.

The history of content moderation is replete with examples of both these weaknesses. Content moderation rules have often been updated haphazardly and in reaction to particular scandals.[104] While public outrage can force an issue, relying on this as a mechanism for policy revision favors those in power and makes platforms highly reactive to a "relatively narrow spectrum of Western media and politics."[105] Judicial-style oversight can make retroactive review of how policies apply in practice more broadly available, and not just to those who can capture public attention. A famous example is Facebook's censorship of the iconic photo known informally as "Napalm Girl."[106] Facebook would not have anticipated this photo being swept up by its rules on nudity or child pornography, and when there was public outrage over the company's takedown of the historic image, Facebook admitted that it had made a mistake and would rewrite its rules to prevent it happening again. But the photo had likely been removed thousands of times before the uproar that led to this admission. It was only

when the person censored was a well-known author whose cause attracted the attention of the press and politicians that policies were revised. Had an internal oversight mechanism been available, the change might have occurred earlier.

Beyond identifying weaknesses in policies, an appeals system can help highlight areas where the policies themselves are fine but the "laws on the books" do not accurately reflect the rules in practice. In any complex system, enforcement error is inevitable. An appeals mechanism acknowledges this reality and formalizes a system for drawing attention to mistakes.

Another benefit of the availability of public review of policies is that it disciplines initial policy makers who know that their decisions may later be subject to public scrutiny and require justification.[107] This does not mean opening up all internal deliberations to public view; decision makers need a measure of opacity to ensure candor and the ability to consult widely.[108] But it does mean that final decisions are more likely to be made based on publicly acceptable grounds because the people who write the rules will be forced to think in advance about whether and how their choices can be rationalized. When rules are later challenged, policy makers are forced to give their reasons and respond to decisions made by an oversight body. By opening up the reasoning process, this public deliberation creates legitimacy for rules the community is asked to abide by.[109]

Provide Public Reasoning

Even where self-regulation does not result in substantively different policies, creating a forum for public contestation and explanation of rules may give them greater legitimacy. "Legitimacy" is a vague concept that can be difficult to define.[110] What is critically important for present purposes is that legitimacy does *not* mean correctness; instead, "in circumstances of relatively widespread reasonable disagreement, . . . legitimacy connote[s] respect-worthiness."[111] This framing helps highlight the very low bar of the legitimacy of the status quo: because content moderation decisions to date have been highly reactive, inconsistent, and poorly justified, they are viewed with little respect by those affected, by civil society, and increasingly by the media and the public at large.

The special nature of speech regulation, where many questions have no "right" answer and any outcome will make certain segments of society feel aggrieved, makes public reasoning an especially important part of garnering greater legitimacy for these decisions. As John Rawls argued, in a pluralistic society where there will always be disagreement about what rule is best, the exercise of power over those who disagree with decisions is *only* legitimized through public reasoning that proceeds in a way people might be expected to respect.[112] The goal is not necessarily to create unanimous agreement, but simply to allow for reasoned disagreement. Empirical research substantiates Rawls's argument: people's judgments of legitimacy do not depend primarily on their obtaining favorable outcomes but are more strongly influenced by the processes and procedures authorities use, including whether they

afford participation, demonstrate impartiality, and show respect for people's interests as worthy of consideration.[113]

This was the case for David Neiwert, whose Twitter account was suspended when he changed his profile picture to the cover of his book about the alt-right, which included Ku Klux Klan hoods.[114] Neiwert thought the suspension was wrong and refused to change the picture; the image was about analyzing hate, not promoting it. Representatives from Twitter reached out and explained that the company takes a no-tolerance stance on such images in profile pictures because they are more prominently displayed on the site. Neiwert wrote that the conversation "was cooperative and [they were] genuinely interested in my input. These Twitter officials were able to persuade me, at least, that they very much share my concerns."[115] Neiwert still disagreed with the decision, but once he understood the reasoning he agreed to change his profile picture.

An independent oversight body can more formally facilitate this process. While it may only be able to do so in a small fraction of cases, as is the case for apex bodies in other legal systems, individual disputes can create opportunities for the articulation of underlying rationales for rules that then flow through to their application in other cases. Indeed, this appears to be an animating reason for the establishment of Facebook's Oversight Board. Facebook already consults with subject-matter experts in the formulation of its community standards,[116] and so it is the public nature of the reasoning process of the Oversight Board that is its distinct offering. Giving public reasons creates decisional transparency and can facilitate a "global dialogue" between a platform and its users about the impacts and justifications of a platform's rules.[117] This responds to criticisms that platforms' opaque decision making interferes with their obligations of clarity, specificity, and predictability.[118] Showing that platforms appreciate the "intellectual, logistic, and moral depth" of content moderation decisions makes their conclusions more accessible and comprehensible.[119]

This can be made more concrete through the example of right-wing provocateur and conspiracy theorist Alex Jones's presence on Facebook. Jones's social-media presence has been a source of long-running controversy. He often spreads highly offensive and inflammatory content as well as dangerous conspiracy theories. But asking platforms that are not democratically accountable to become "arbiters of truth" and engage in drawing hard lines about whether commentary is false or acceptable implicates free speech concerns. In response to one particular post by Jones, leaked emails show that internal executives were having difficulty deciding whether the post violated Facebook's Community Standards.[120] One executive referred to the fact that the number of violating comments on the post "d[id] not meet the threshold for deletion."[121] UK executives then pointed to local context, noting that the image in question is famous in the United Kingdom, where it is "widely acknowledged to be anti-Semitic."[122] The idea of a "threshold" is not explicit in Facebook's Community Standards, and those not familiar with the UK context did not understand the

full import of the image. Without public explanation, a decision either way on the basis of these considerations would be opaque and hard to evaluate. It would provide little guidance for users about what behavior does or does not violate Facebook's policies. Ventilating these arguments for and against deletion through public oversight would help legitimize them. The response to the eventual removal of Jones from Facebook altogether in May 2019 substantiates this. Although many commentators had been calling for his removal for some time, the response was mixed. Few defended Jones, but there was frustration with the way Facebook executed the ban and the lack of transparency around the reason or timing.[123] Without public reasoning, even those who agreed with the decision thought it was illegitimate.

It's important to reiterate: public reasoning cannot create substantive agreement with all decisions. The matters involved will always be controversial. But because of the public significance of speech rules, public reasoning is the *only* way to make these decisions accountable and increase their legitimacy. In the short term, the dividends may not always be obvious. When a doctored video of House Speaker Nancy Pelosi, which had been slowed down to make her appear drunk, started spreading on social media, Facebook left the video up, and a high-level executive appeared on national television to explain the company's reasoning. YouTube meanwhile quietly took the video down and gave a dubious rationale only when asked.[124] Yet it was still Facebook, not YouTube, that faced the brunt of public outrage. But concerns about legitimacy are concerns about the longer-term functioning of the overall system, not only individual cases. A few weeks later, YouTube was enveloped in a storm of controversy because of the way it handled complaints from a journalist who was being harassed by another YouTube creator. YouTube responded mostly via Twitter in abrupt tweets and seemingly kept reversing its position. No one involved was happy with the final resolution, and there was public outrage from all sides at YouTube's clumsy handling of the matter.[125] Demands that platforms become publicly accountable for the way they exercise their power in matters of public importance are, at heart, calls for public reasoning, and this is what an oversight body can provide.

Develop and Explain Norms

Another consequential benefit of public reasoning, beyond legitimation, is the greater visibility brought to community norms. Through contestation and explanation of these norms in a more public forum, more community members may become aware of the rules, which in turn helps generate compliance. James Grimmelmann describes the importance of this aspect, saying that moderation's "most important mission is to create strong shared norms among participants. . . . Moderators can influence norms directly by articulating them. They can do this either in general, with codes of conduct and other broad statements of rules, or in specific cases by praising good behavior and criticizing bad."[126] J. Nathan Matias similarly found that the visibility of the rules of online communities substantially increases compliance and overall participation in the community.[127] By allowing greater visibility and participation in content moderation decisions through public reasoning,

internal oversight can embed the process of rule formation in a broader community and help norms be formed and tested.

Diffuse Pressure

For platforms, an independent oversight mechanism might provide a way to outsource controversial decisions.[128] But this may also be beneficial for other stakeholders. One of the reasons why it is especially important who decides what speech is acceptable is the fear, and historical actuality, of the use of government or majority power to silence unpopular voices. As described above, if governments seek to use their power to silence unpopular speech, in most cases a platform oversight body will offer little protection. But platform oversight can help protect minorities against majorities that seek to pressure a platform directly and not through governments. A meaningfully independent oversight body need not be responsive to majoritarian will or commercial pressure. It can therefore be an important protection for otherwise marginalized voices in the public sphere. Of course, the extent to which this protective aspiration is realized depends on the specifics of the institutional design and on ensuring that those charged with oversight are properly incentivized and not susceptible to pressure.[129]

Explain Government Moderation

When it comes to the relationship between platforms and governments, the focus has generally been on fears that governments will use platforms to achieve speech regulation that they themselves could not lawfully do directly. Jack M. Balkin calls this "collateral censorship."[130] As discussed above, collateral censorship is one content moderation concern that self-regulatory oversight cannot meaningfully address.[131]

By contrast, governments increasingly also find themselves on the other side of content moderation. As social media platforms become important forums for governments and politicians to communicate with their constituents, questions more frequently arise about how platforms should treat content from these actors that violates their rules.[132] This is one of the thorniest issues in content moderation. If platforms lack democratic accountability or legitimacy, this concern is strongest when a platform interferes with communications between a government and its polity. This might be especially true when the government is democratic, but arguably citizens have an interest in knowing the views or actions of their rulers even when they did not vote for them. At the very least, it is problematic for a private company (and in most jurisdictions, a *foreign* private company) to decide that the citizens have no such entitlement.

There are no easy ways to make these decisions about how to balance the competing interests that arise in these cases—for example, when a politician uses hateful language, the prevention of harassment or hate speech and the public interest in relevant information come into conflict.[133] But platform decisions in these cases have, to date, been ad hoc and often reflective of real-world power dynamics.

For example, the suggestion has been made, only through leaks to the media, that a reason Twitter has been slow to remove white-supremacist content from its platform is because it might sweep up Republican politicians. This perception delegitimizes Twitter's decision not to adopt a removal policy because it becomes viewed as arbitrary or biased.[134] Similarly, it was only through comments to the press that Facebook announced it officially had a "state actor" policy that exempted governments from the usual rules in its community standards.[135] This tactic also was to the detriment of the perceived legitimacy of that decision. The policy is not reflected in its public-facing standards and is not consistently applied; politicians in less important markets (for example, Australia[136]) have been censored, while those in the United States have not. The opacity and malleability of these rules is problematic in a context where Facebook's content moderation decisions have such direct impact on the relationship between people and their governments.

Clear rules alone are not enough. To deal with some of the above problems, Twitter recently announced an updated policy about how it will treat tweets by state actors that breach its rules but whose availability might be in the public interest.[137] The new policy includes a set of detailed criteria that will be used to assess whether a particular tweet should be taken down. But their application will not always be straightforward: the "immediacy and severity of potential harm" of a threat of nuclear war by a government leader, for example, is debatable. Twitter's decisions under this policy will, by definition, be political and no doubt contested. Therefore, public reasoning is necessary to bring clarity and coherence to these decisions.

The role of rationalizing and legitimizing how platforms treat government speech is only a subset of the role that an oversight body has in moderating content more generally. But as platforms are increasingly both the weapon of and a form of exerting power over many governments, this is an especially important function. It is also a pertinent reminder of the special difficulties inherent in regulating content moderation given that governments' own interests are at stake.

Conclusion

The intricacies of content moderation require enforcers to be close to the ground but within a larger ecosystem that makes enforcement scalable.[138] Governments cannot and should not perform this role for speech. Practically, translating broad standards into individual decisions would stretch government capacity given the volume at which these platforms operate. Normatively, such extensive government interference with speech regulation is inconsistent with the underlying purpose of the right to free speech. But current content moderation practices are suffering from a crisis of trust. Therefore, platforms need to develop more robust internal oversight mechanisms to bring greater legitimacy and coherence to their speech decisions.

Currently, highly consequential decisions about free speech rights are playing out in settings that are not well suited to them, whether they be behind closed doors at private

companies, in the midst of public outrage covered by the media, or in ways that do not adequately reflect the very particular relationship that governments should have with speech regulation. Individuals and the public interest are poorly represented in these conversations. The format does not help foster reasoned disagreement. Internal oversight mechanisms, both required and legitimized by government regulation, can help create forums that deal with these difficult issues more productively and in ways that produce greater accountability. Work on designing the forms of these systems is only just beginning. It needs to start by being realistic about what regulation can and can't achieve, which allows a more conscious evaluation of trade-offs to take place. A model of "verified accountability" will have shortcomings; it may only be the worst system except for all the rest. But it is a beta version of a new form of platform governance that shows promise.

NOTES

1 Nicolas Suzor, "Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms," *Social Media + Society* 4, no. 3 (2018): 1–11, 1.

2 Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven, CT: Yale University Press, 2018), 5.

3 Andrew Keane Woods, *Tech Firms Are Not Sovereigns*, Aegis Series Paper No. 1813 (Stanford, CA: Hoover Institution, 2018); Kristen E. Eichensehr, "Digital Switzerlands," *University of Pennsylvania Law Review* 167 (2019, forthcoming).

4 Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634; Facebook, "Draft Charter for Oversight Board for Content Decisions," January 28, 2019, https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-1.pdf, archived at https://perma.cc/5C99-C9JF.

5 Zoe Mentel Darmé, Matt Miller, and Kevin Steeves, "Global Feedback & Input on the Facebook Oversight Board for Content Decisions," Facebook, June 27, 2019, https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-1.pdf, 37.

6 Decisions about speech on social media impact vast areas of modern life, from the spread of political content during elections to the spread of hate speech in volatile societies: see, e.g., Nathaniel Persily, "The Internet's Challenge to Democracy: Framing the Problem and Assessing Reforms," Kofi Annan Foundation, 2019, https://pacscenter.stanford.edu/publication/the-internets-challenge-to-democracy-framing-the-problem-and-assessing-reforms.

7 Monika Bickert, "Defining the Boundaries of Free Speech on Social Media," in *The Free Speech Century*, edited by Lee C. Bollinger and Geoffrey R. Stone (New York: Oxford University Press, 2018), 254, 257.

8 French Secretary of State for Digital Affairs, "Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision," May 2019, https://minefi.hosting.augure.com/Augure_Minefi/r/ContenuEnLigne/Download?id=AE5B7ED5-2385-4749-9CE8-E4E1B36873E4&filename=Mission%20Re%CC%81gulation%20des%20re%CC%81seaux%20sociaux%20-ENG.pdf.

9 As Fallon notes, "legitimacy" itself is a word with many meanings: Richard H. Fallon Jr., *Law and Legitimacy in the Supreme Court* (Cambridge, MA: The Belknap Press of Harvard University Press, 2018), 6.

10 See, e.g., "Beware of Tech Companies Playing Government," January 17, 2019, https://www.bloomberg.com/opinion/articles/2019-01-17/beware-of-tech-companies-playing-government; Cindy Cohn, "Social Media Platforms Should be Accountable and Transparent About Content Removal, But DOJ's Plan to Investigate Raises Concerns," Electronic Frontier Foundation, September 6, 2018, https://www.eff.org/deeplinks/2018/09/whats-doj-really-seeking-accountability-content-moderation-or-censoring-speech-it; "CDT, Coalition Urge Internet Platforms to Provide More Transparency and Accountability in Content Moderation," Center for Democracy & Technology, https://cdt.org/press/cdt-coalition-urge-internet-platforms-to-provide-more-transparency-and-accountability-in-content-moderation.

11 See, e.g., Timothy Garton Ash, Robert Gorwa, and Danaë Metaxa, *GLASNOST! Nine Ways Facebook Can Make Itself a Better Forum for Free Speech and Democracy* (Oxford: Reuters Institute and Oxford University Press, 2019), https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Garton_Ash_et_al_Facebook_report_FINAL_0.pdf, 18; David Kaye, "Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression," United Nations Human Rights Council, June 4, 2018, https://digitallibrary.un.org/record/1637534; Nicolas Suzor, Tess van Geelen, and Sarah Myers West, "Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda," *International Communication Gazette* 80, no. 4 (2018): 385–400.

12 Bickert, "Defining the Boundaries of Free Speech," 265.

13 Facebook Data Transparency Advisory Group, "Report of The Facebook Data Transparency Advisory Group," Justice Collaboratory, Yale Law School, April 2019, https://law.yale.edu/system/files/area/center/justice/document/dtag_report_5.22.2019.pdf, 39.

14 Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review* 131 (2018): 1598.

15 Thomas Kadri and Kate Klonick, "Facebook v. Sullivan: Building Constitutional Law for Online Speech," *Southern California Law Review* (forthcoming).

16 See, e.g., Martha Minow, "Alternatives to the State Action Doctrine in the Era of Privatization, Mandatory Arbitration, and the Internet: Directing Law to Serve Human Needs," *Harvard Civil Rights–Civil Liberties Law Review* 52, no. 1 (2017): 145–67.

17 Jody Freeman, "The Private Role in the Public Governance," *New York University Law Review* 75, no. 3 (2000): 543, 576.

18 Freeman, "The Private Role in the Public Governance"; Minow, "Alternatives to the State Action Doctrine."

19 Manhattan Community Access Corp. v. Halleck, 587 U.S. ___, 10 (2019). The dissent also appeared to show little sympathy for the view that private platforms might be constrained by the First Amendment, but it differed with the majority on the facts of the particular case, which involved a television cable channel.

20 John Perry Barlow, "A Declaration of the Independence of Cyberspace," Electronic Frontier Foundation, February 8, 1996, https://www.eff.org/cyberspace-independence.

21 Mark Tushnet, "The Issue of State Action/Horizontal Effect in Comparative Constitutional Law," *International Journal of Constitutional Law* 1, no. 1 (2003): 79, 89.

22 Beth Stephens, "The Amorality of Profit: Transnational Corporations and Human Rights," *Berkeley Journal of International Law* 20, no. 1 (2002): 62–63; Monika Zalnieriute, "From Human Rights Aspirations to Enforceable Obligations by Non-State Actors in the Digital Age: The Example of Internet Governance and ICANN," *Yale Journal of Law and Technology* 24 (2019, forthcoming).

23 Tushnet, "The Issue of State Action."

24 United Nations Human Rights Council, "UN Guiding Principles on Business and Human Rights (A/HRC/17/31)," 2011, Principle 11, https://www.ohchr.org/Documents/Issues/Business/A-HRC-17-31_AEV.pdf.

25 Woods, *Tech Firms Are Not Sovereigns*; Frank Pasquale, "From Territorial to Functional Sovereignty: The Case of Amazon," *Law and Political Economy*, December 6, 2017, https://lpeblog.org/2017/12/06/from-territorial-to-functional-sovereignty-the-case-of-amazon.

26 K. Sabeel Rahman, "The New Utilities: Private Power, Social Infrastructure, and the Revival of the Public Utility Concept," *Cardozo Law Review* 39 (2018): 1621–89.

27 Packingham v. North Carolina, 582 U.S. __ (2017).

28 Heather Whitney, "Search Engines, Social Media, and the Editorial Analogy," Knight First Amendment Institute Emerging Threats Series, 2018, https://knightcolumbia.org/sites/default/files/content/Heather_Whitney_Search_Engines_Editorial_Analogy.pdf, 24.

29 Mark Bunting, "From Editorial Obligation to Procedural Accountability: Policy Approaches to Online Content in the Era of Information Intermediaries," *Journal of Cyber Policy* 3, no. 2 (2018): 165, 169.

30 Jack M. Balkin, "Fixing Social Media's Grand Bargain," Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1814, October 15, 2018, https://www.hoover.org/research/fixing-social-medias-grand-bargain, 6–7.

31 David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (New York: Columbia Global Reports, 2019) 17, 52.

32 Zalnieriute, "From Human Rights Aspirations to Enforceable Obligations."

33 Daphne Keller, "Internet Platforms: Observations on Speech, Danger, and Money," Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1807, June 13, 2018, https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf, 2.

34 Daphne Keller, "The Right Tools: Europe's Intermediary Liability Laws and the 2016 General Data Protection Regulation," *Berkeley Technology Law Journal* 33, no. 1 (2018): 287–364.

35 See, e.g., similar discussion in Eichensehr, "Digital Switzerlands." See also Evelyn Douek, "YouTube's Bad Week and the Limitations of Laboratories of Online Governance," *Lawfare*, June 11, 2019, https://www.lawfareblog.com/youtubes-bad-week-and-limitations-laboratories-online-governance.

36 French Secretary of State for Digital Affairs, "Creating a French Framework."

37 For more on this relationship between voluntary initiatives and government regulation, see Emily B. Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Social Responsibility* (Cambridge: Cambridge University Press, 2015), 78–83.

38 Frederick Schauer, "Fear, Risk and the First Amendment: Unraveling the Chilling Effect," *Boston University Law Review* 58 (1978): 691–92.

39 See, e.g., Laidlaw, *Regulating Speech in Cyberspace*, 45.

40 Schauer, "Fear, Risk and the First Amendment," 685.

41 Sarah H. Cleveland, "Hate Speech at Home and Abroad," in *The Free Speech Century*, edited by Lee C. Bollinger and Geoffrey R. Stone (New York: Oxford University Press, 2018), 210, 224.

42 See, e.g., Ronald J. Krotoszynski Jr., *The First Amendment in Cross-Cultural Perspective: A Comparative Legal Analysis of the Freedom of Speech* (New York: New York University Press, 2009); Michel Rosenfeld, "Hate Speech in Constitutional Jurisprudence: A Comparative Analysis Conference: The Inaugural Conference of the Floersheimer Center for Constitutional Democracy: Fundamentalisms, Equalities, and the Challenge to Tolerance in a Post-9/11 Environment," *Cardozo Law Review* 24, no. 4 (2003), 1523–67.

43 Ash et al., *GLASNOST!*, 11–12.

44 Bickert, "Defining the Boundaries of Free Speech," 257.

45 Mark Tushnet, "How Different Are Waldron's and Fallon's Core Cases For and Against Judicial Review?," *Oxford Journal of Legal Studies* 30, no. 1 (2010): 55.

46 See, e.g., the explanation of the conditions that must be met for limitations on freedom of expression under international law in Kaye, "Report of the Special Rapporteur," 4–5 ¶7.

47 See, e.g., International Covenant on Civil and Political Rights, art. 19(3)b.

48 Gillespie, *Custodians of the Internet*, 13.

49 Sarah Mervosh, "'Knitting Has Always Been Political': Ravelry Bans Pro-Trump Content, and Reactions Flood In," *New York Times*, June 24, 2019, https://www.nytimes.com/2019/06/24/style/ravelry-knitting-ban-trump.html.

50 Frederick Schauer, "The Exceptional First Amendment," KSG Working Paper No. RWP05-021, February 2005, http://dx.doi.org/10.2139/ssrn.668543, 24; Elena Kagan, "Private Speech, Public Purpose: The Role of Governmental Motive in First Amendment Doctrine," *University of Chicago Law Review* 63, no. 2 (1996): 413–517.

51 French Secretary of State for Digital Affairs, "Creating a French Framework," 13.

52 Cass R Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton, NJ: Princeton University Press, 2017), 205.

53 Kaye, *Speech Police*, 126 ("Ultimately, we need to answer the question—who is to be in charge?—in a way that works for us, as a public and as individuals, that enables us to claw back some part of the original promise of democratic space the internet originally offered.").

54 Berkman Klein Center, *How to Work with Tech Companies on Human Rights*, Berkman Klein Center, April 25–28, 2019 (David Sullivan, of the Global Network Initiative, comparing secrecy in multi-stakeholder initiatives around conflict minerals with the GNI), https://cyber.harvard.edu/events/2019-04-23/how-work-tech-companies-human-rights. For critiques of the GNI's secrecy, see Laidlaw, *Regulating Speech in Cyberspace*, 104–107; Rikke Frank Jørgensen, "Human Rights and Private Actors in the Online Domain," in *New Technology for Human Rights Law and Practice,* edited by Molly K. Land and Jay D. Aronson (Cambridge: Cambridge University Press, 2018), 243–69.

55 Rory Van Loo, "The Corporation as Courthouse," *Yale Journal on Regulation* 33, no. 2 (2016): 554; Cf. Josh Dzieza, "Dirty Dealing in the $175 Billion Amazon Marketplace," *The Verge*, December 19, 2018, https://www.theverge.com/2018/12/19/18140799/amazon-marketplace-scams-seller-court-appeal-reinstatement.

56 Ethan Katsh and Orna Rabinovich-Einy, *Digital Justice: Technology and the Internet of Disputes* (New York: Oxford University Press, 2017), 38.

57 Katsh and Rabinovich-Einy, *Digital Justice*, 34, 114; Mary Anne Franks, "Justice Beyond Dispute—Book Review: *Digital Justice,*" *Harvard Law Review* 131 (2018), 1378–79.

58 For present purposes, I have set aside the issue of determining which platforms government regulation should apply to. This is an important question, and government regulation must be careful not to impose burdens that smaller platforms do not have the resources to comply with, thereby entrenching the dominant platforms. The discussion that follows focuses on the major and systemically important platforms, accepting that this category may be hard to define in practice.

59 French Secretary of State for Digital Affairs, "Creating a French Framework," 11.

60 See, e.g., ACLU Legal Foundation of Northern California et al., "The Santa Clara Principles: On Transparency and Accountability in Content Moderation," https://santaclaraprinciples.org; Kaye, "Report of the Special Rapporteur"; Ash et al., *GLASNOST!*; Danielle Keats Citron, "Extremist Speech, Compelled Conformity, and Censorship Creep," *Notre Dame Law Review* 93, no. 3 (2018): 1035–71.

61 Yuval Eylon and Alon Harel, "The Right to Judicial Review," *Virginia Law Review* 92, no. 5 (2006): 997.

62 Alex Stamos, "2018 CISAC Drell Lecture: The Battle for the Soul of the Internet," 2018, https://www.youtube.com/watch?v=NKN6xLhTjIo&feature=youtu.be&t=1222.

63 Facebook Data Transparency Advisory Group, "Report of The Facebook Data Transparency Advisory Group," 34.

64 Freedman v. Maryland, 380 U.S. 51, 58 (1965); Henry P. Monaghan, "First Amendment Due Process," *Harvard Law Review* 83, no. 3 (1970): 519.

65 Morrissey v. Brewer, 408 U.S. 471, 481 (1972). See also Adrian Vermeule, "Deference and Due Process," *Harvard Law Review* 129, no. 7 (2016): 1896.

66 Mathews v. Eldridge, 424 U.S. 319 (1976).

67 Vermeule, "Deference and Due Process," 1903–4.

68 Walters v. National Association of Radiation Survivors, 473 U.S. 305, 321 (1985), quoting Mathews, 424 U.S. 319, 344 (1976).

69 ACLU Legal Foundation of Northern California et al., "The Santa Clara Principles."

70 The data in this paragraph come from Facebook, "Community Standards Enforcement," May 23, 2019, https://transparency.facebook.com/community-standards-enforcement.

71 Zuckerberg, "A Blueprint for Content Governance and Enforcement."

72 I am not arguing that Facebook could not afford to devote more resources to providing affected users with greater procedural protections. It certainly could. But the scale of the platform means that even a company as profitable as Facebook will ultimately be forced to make trade-offs in the process it affords. Two *million* people were estimated to be employed in efforts to monitor and censor the internet in China during the year 2013, in a system that provides far less process and accuracy (two factors that increase the resource intensity of moderation) than we might expect in Western countries: Elizabeth C. Economy, "The Great Firewall of China: Xi Jinping's Internet Shutdown," *The Guardian*, June 29, 2018, https://www.theguardian.com/news/2018/jun/29/the-great-firewall-of-china-xi-jinpings-internet-shutdown.

73 Lawrence B. Solum, "Procedural Justice," *Southern California Law Review* 78, no. 1 (2004): 251–52 (2004).

74 See, e.g., Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM), 15 U.S.C §§ 7701–7713 (2003).

75 As noted in the introduction, self-regulatory appeals bodies can only be effective if embedded in an entire system of governance. The form of this larger system is beyond the scope of this paper.

76 Kaye, "Report of the Special Rapporteur," 18.

77 Facebook, "Community Standards Enforcement."

78 Elizabeth Dwoskin, "How YouTube Erased History in Its Battle against White Supremacy," *Washington Post*, June 13, 2019, https://www.washingtonpost.com/technology/2019/06/13/how-youtube-erased-history-its-battle-against-white-supremacy (describing how YouTube accidentally removed educational and journalistic videos when launching a new policy on white supremacists and hoaxes).

79 "Facebook Says First-Person Christchurch Video Foiled AI System," April 24, 2019, https://www.bloomberg.com/news/articles/2019-04-24/facebook-says-first-person-christchurch-video-foiled-ai-system; "A Further Update on New Zealand Terrorist Attack," Facebook Newsroom, March 20, 2019 https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand.

80 Gillespie, *Custodians of the Internet*, 9.

81 Tim Wu, "Is the First Amendment Obsolete?," Knight First Amendment Institute's Emerging Threats, 2017, https://knightcolumbia.org/sites/default/files/content/Emerging%20Threats%20Tim%20Wu%20Is%20the%20First%20Amendment%20Obsolete.pdf, 7.

82 Zeynep Tufekci, "It's the (Democracy-Poisoning) Golden Age of Free Speech," *WIRED*, January 18, 2018, https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship.

83 Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York, "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation," *International Journal of Communication* 13 (2019): 1527.

84 For discussion of the pitfalls of what he calls the "quasi-religious significance" assigned to transparency, see David E. Pozen, "Transparency's Ideological Drift," *Yale Law Journal* 128, no. 1 (2018): 161.

85 Cass R. Sunstein, "Output Transparency vs. Input Transparency," May 25, 2017, https://ssrn.com/abstract=2826009.

86 Lillian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," *Duke Law & Technology Review* 16, no. 1 (2017): 67.

87 Facebook, "Community Standards Enforcement."

88 French Secretary of State for Digital Affairs, "Creating a French Framework," 12.

89 Suzor et al., "What Do We Mean When We Talk About Transparency?," 1529.

90 Kaye, "Report of the Special Rapporteur," 20.

91 Daphne Keller, "The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation," *Berkeley Technology Law Journal* 33, no. 1 (2018): 340.

92 Laura Murphy, "Facebook's Civil Rights Audit," June 30, 2019, https://fbnewsroomus.files.wordpress.com/2019/06/civilrightaudit_final.pdf.

93 Bickert, "Defining the Boundaries of Free Speech," 262.

94 Facebook, "Draft Charter for Oversight Board for Content Decisions," 1–2.

95 Ash et al., *GLASNOST!*

96 Klonick, "The New Governors," 1663–64.

97 Ash et al., *GLASNOST!,* 11–12.

98 Kaye, "Report of the Special Rapporteur," at ¶42.

99 The following section draws on Evelyn Douek, "Facebook's 'Oversight Board': Move Fast with Stable Infrastructure and Humility," *North Carolina Journal of Law & Technology* 21, no. 1 (2019, forthcoming).

100 For more discussion of this proposal specifically, see Douek, "Facebook's 'Oversight Board.'"

101 Rosalind Dixon, "The Core Case for Weak-Form Judicial Review," *Cardozo Law Review* 38, no. 6 (2017): 2193–2232.

102 Dixon, "The Core Case for Weak-Form Judicial Review," 2208–9.

103 Dixon, "The Core Case for Weak-Form Judicial Review," 2209–12.

104 See, e.g., Gillespie, *Custodians of the Internet*, 65–66; Klonick, "The New Governors," 1630–35, 1648–58; Ash et al., *GLASNOST!*, 9.

105 Ash et al., *GLASNOST!*, 9.

106 Gillespie, *Custodians of the Internet*, 1–4.

107 Alec Stone Sweet, *Governing with Judges: Constitutional Politics in Europe* (Oxford: Oxford University Press, 2000) 62, 74–75; Juliane Kokott and Martin Kaspar, "Ensuring Constitutional Efficacy," in *The Oxford Handbook of Comparative Constitutional Law* (Oxford: Oxford University Press, 2012), 806.

108 Sunstein, "Output Transparency vs. Input Transparency," 3.

109 Dixon, "The Core Case for Weak-Form Judicial Review," 2224, citing Cass R. Sunstein, *One Case at a Time: Judicial Minimalism on the Supreme Court* (Cambridge, MA: Harvard University Press, 1999).

110 Fallon, *Law and Legitimacy*, 6.

111 Fallon, *Law and Legitimacy*, 8.

112 John Rawls, *Political Liberalism*, expanded ed. (New York: Columbia University Press, 2005), 217; see also Fallon, *Law and Legitimacy*, 12; Solum, "Procedural Justice," 230–31.

113 Tom R. Tyler, "Procedural Justice, Legitimacy, and the Effective Rule of Law," *Crime and Justice* 30 (2003): 283–357; Facebook Data Transparency Advisory Group, "Report of The Facebook Data Transparency Advisory Group," 34.

114 David Neiwert, "Why I Ended My Fight with Twitter—For Now," *Daily Kos*, June 25, 2019, https://www.dailykos.com/story/2019/6/25/1867213/-Why-I-ended-my-fight-with-Twitter-for-now.

115 Neiwert, "Why I Ended My Fight with Twitter."

116 "Facts About Content Review on Facebook," Facebook Newsroom, December 28, 2018, https://newsroom.fb.com/news/2018/12/content-review-facts.

117 Ash et al., *GLASNOST!*, 13; Kaye, "Report of the Special Rapporteur," 19.

118 Kaye, "Report of the Special Rapporteur," 19.

119 Facebook Data Transparency Advisory Group, "Report of The Facebook Data Transparency Advisory Group," 41.

120 Jake Kanter, "Leaked Emails Reveal Facebook's Intense Internal Discussion over Alex Jones' 'Anti-Semitic' Post on Instagram," Business Insider, https://www.businessinsider.com/facebook-emails-reveal-discussion-about-alex-jones-instagram-account-2019-3.

121 Kanter, "Leaked Emails Reveal Facebook's Intense Internal Discussion." Infringing comments amounted to about four percent of the total number of comments on the post.

122 Kanter, "Leaked Emails Reveal Facebook's Intense Internal Discussion."

123 April Glaser, "Why Facebook's Latest Ban of Alex Jones and Company Was So Underwhelming," *Slate Magazine* ("Facebook has the power to punish wrongdoers, as it did on Thursday. But we don't know its full rationale for doing so, nor do we know who will be next."), May 3, 2019, https://slate.com/technology/2019/05/facebook-alex-jones-ban-underwhelming.html; Bret Stephens, "Opinion: Facebook's Unintended Consequence," *New York Times*, May 3, 2019 ("The deeper problem is the overwhelming concentration of technical, financial and moral power in the hands of people who lack the training, experience, wisdom, trustworthiness, humility and incentives to exercise that power responsibly. . . . The decision to absolutely ban certain individuals will always be a human one. It will inevitably be subjective."), https://www.nytimes.com/2019/05/03/opinion/facebook-free-speech.html; "Why Facebook's Bans Warrant Concern," *National Review* ("This means a person can potentially face social-media bans even if they comply with every syllable of the company's speech rules on the company's platform. The potential for abuse is obvious, as is the potential chilling effect."), May 3, 2019, https://www.nationalreview.com/corner/why-facebooks-bans-warrant-concern; Emily Stewart, "Facebook Bans Alex Jones, Infowars, Louis Farrakhan, and Others It Deems 'Dangerous,'" *Vox* ("It's not clear why Facebook is doing this now, but pressure for it to take action has been mounting for quite some time, and the decision is probably at least in part an effort to get some positive PR."), May 2, 2019, https://www.vox.com/recode/2019/5/2/18527357/facebook-bans-alex-jones-louis-farrakhan-infowars.

124 Douek, "YouTube's Bad Week."

125 Douek, "YouTube's Bad Week."

126 James Grimmelmann, "The Virtues of Moderation," *Yale Journal of Law & Technology* 17 (2015): 61–62.

127 J. Nathan Matias, "Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions," *Proceedings of the National Academy of Sciences of the United States of America* 116 (2019): 9785.

128 See Douek, "Facebook's 'Oversight Board.'"

129 The best ways to achieve this are beyond the ambit of this paper, but the need for self-regulation to include some element that is both independent as well as *seen* to be a meaningful independent check on business decisions is critical to its success.

130 Jack M. Balkin, "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation," *University of California, Davis Law Review* 51 (2018): 1175–77.

131 See the "Global Norms" section in this essay.

132 See, e.g., Kadri and Klonick, "Facebook v. Sullivan," 33–35; Ryan Mac, "Twitter Just Removed A Tweet From An Account Linked To Iran's Supreme Leader, Raising Enforcement Questions," BuzzFeed News, February 15, 2019, https://www.buzzfeednews.com/article/ryanmac/twitter-removes-tweet-reportedly-from-irans-supreme-leader; Jake Evans, "Fraser Anning's Public Facebook Page Removed," ABC News, September 28, 2018, https://www.abc.net.au/news/2018-09-28/fraser-annings-facebook-page-taken-down/10317638.

133 Kadri and Klonick, "Facebook v. Sullivan," 18–21.

134 Joseph Cox and Jason Koebler, "Why Won't Twitter Treat White Supremacy Like ISIS? Because Would Mean Banning Some Republican Politicians Too," Vice, April 25, 2019, https://www.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too.

135 David Ingram, "Facebook's New Rapid Response Team Has a Crucial Task: Avoid Fueling Another Genocide, NBC News, June 20, 2019, https://www.nbcnews.com/tech/tech-news/facebook-s-new-rapid-response-team-has-crucial-task-avoid-n1019821.

136 See, e.g., Evans, "Fraser Anning's Public Facebook Page Removed."

137 Twitter Safety, "Defining Public Interest on Twitter," Twitter, June 27, 2019, https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html.

138 Bickert, "Defining the Boundaries of Free Speech," 257.

25 24 23 22 21 20 19 7 6 5 4 3 2 1

The preferred citation for this publication is:
Evelyn Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation,* Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1903, available at https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-regulation

About the Author

EVELYN DOUEK

Evelyn Douek is an SJD candidate at Harvard Law School and an affiliate at the Berkman Klein Center for Internet & Society. She studies international and transnational regulation of online speech and content-moderation institutional design. Prior to coming to Harvard, Douek was an associate (clerk) to the Honourable Chief Justice Susan Kiefel of the High Court of Australia and worked in commercial litigation in Sydney.

Working Group on National Security, Technology, and Law

The Working Group on National Security, Technology, and Law brings together national and international specialists with broad interdisciplinary expertise to analyze how technology affects national security and national security law and how governments can use that technology to defend themselves, consistent with constitutional values and the rule of law.

The group focuses on a broad range of interests, from surveillance to counterterrorism to the dramatic impact that rapid technological change—digitalization, computerization, miniaturization, and automaticity—are having on national security and national security law. Topics include cybersecurity, the rise of drones and autonomous weapons systems, and the need for—and dangers of—state surveillance. The group's output will also be published on the Lawfare blog, which covers the merits of the underlying legal and policy debates of actions taken or contemplated to protect the nation and the nation's laws and legal institutions.

Jack Goldsmith is the chair of the National Security, Technology, and Law Working Group.

For more information about this Hoover Institution Working Group, visit us online at http://www.hoover.org/research-teams /national-security-technology-law-working-group.

Hoover Institution, Stanford University
434 Galvez Mall
Stanford, CA 94305-6003
650-723-1754

Hoover Institution in Washington
The Johnson Center
1399 New York Avenue NW, Suite 500
Washington, DC 20005
202-760-3200

HOOVER INSTITUTION