

CONTENT OR CONTEXT MODERATION?

*Artisanal, Community-Reliant,
and Industrial Approaches*

ROBYN
CAPLAN

EXECUTIVE SUMMARY

Platform companies are increasingly being called upon to make ethical decisions regarding speech, and in response, the public is becoming more interested in how content moderation policies are formed by platform employees. At the same time, interest in this topic has overwhelmingly focused on major social media platforms, such as Facebook, which is a rare type of organization when considering content moderation. This report investigates the differences *between* platform companies of various sizes and examines the tensions and tradeoffs organizations must make as they set rules and guidelines for content online.

Through an interview study with representatives from 10 major platforms, this report explores the resource gaps, in terms of diversity and amount of personnel and technology, which exist in platform companies of different **missions, business models**, and **size of team**. In particular, it focuses on three different models of content moderation:

- 1) **Artisanal**, for platforms such as Vimeo, Medium, Patreon, or Discord;
- 2) **Community-Reliant**, for platforms such as Wikimedia and Reddit; and
- 3) **Industrial** approaches, for platforms such as Facebook or Google.

As these companies make content policy that has an impact on citizens around the world, they must carefully consider how to be sensitive to localized variations in how issues like context-based speech, like hate speech and disinformation, manifest in different regions and political circumstances. At the same time, due to the scale at which they are operating, these companies are often working to establish consistent rules, both to increase transparency for users and to operationalize their enforcement for employees. This report contends that the three different approaches prioritize this balance between **context-sensitivity** and **consistency** differently, depending on resource needs and organizational dynamics.

Understanding these differences and the nuances of each organization is helpful for determining both the expectations we should be placing on companies and the range of solutions that need to be brought to bear, including existing legislation such as Section 230 of the Communications Decency Act in the United States and the NetzDG rule in Germany. This is important for the artisanal organizations that need to formalize their logic to address concerns more consistently. And it is also important for the industrial-sized operations that need to translate values into training and evaluations while being sensitive to the individual differences of content, such as hate speech and newsworthiness.

TABLE OF CONTENTS

Introduction	3
What are Platforms, According to Whom?	8
Balancing Context Moderation with Consistent Moderation	13
Approaches to Content and Context Moderation: Artisanal, Community-Reliant, and Industrial	15
The Artisanal Approach	17
The Community-Reliant Approach	20
The Industrial Approach	23
Consistent Versus Contextual Platform Regulation	26
Conclusion: Sizing Up Platform Regulation	31
Acknowledgments	35

Robyn Caplan; Affiliate, Data & Society; PhD Candidate, School of Communication and Information Studies, Rutgers University

Supported by a grant from the Open Society Foundations.

INTRODUCTION

In 2012, journalist Adrian Chen published an article describing a workforce most individuals did not know existed: content moderators. “Anti-Porn and Gore Brigade” described the outsourced workers Facebook contracted who operated remotely from Turkey, the Philippines, Mexico, and India to review content flagged as violating their community standards.¹ Chen’s article was a response to then-recent controversies around Facebook’s content moderation. A public outcry over “censorship” had arisen when journalists had uncovered evidence that the company had been removing content that was widely considered acceptable in the US, such as images of women breastfeeding or two men kissing. A lot has happened since then. Concerns about the rise of hate speech and disinformation have increased the amount of public scrutiny being placed on search engine and social media companies that are responsible for mediating much of the world’s information. Part of this scrutiny concerns the lack of transparency into the content moderation rules and a lack of visibility into *how* platforms are developing these rules.² In 2017, journalist Julia Angwin’s piece for ProPublica – “Facebook’s Secret Censorship Rules Protect White Men but Not Black Children” – hit at the heart of the debate, highlighting how tech companies, when made responsible for establishing the difference between hate speech and political expression, often search for *straightforward, consistent calculations*, which are all too often divorced from historical and cultural contexts.³ In 2018, an investigation of content moderation at Facebook by British broadcaster Channel 4 showed the other end of the spectrum. Leaving content moderation decisions largely to the *discretion of individual workers* can lead to rules being applied inconsistently, letting through images of child abuse and instances of hate speech.⁴

Concerns about the rise of hate speech and disinformation have increased the amount of public scrutiny being placed on search engine and social media companies that are responsible for mediating much of the world’s information.

1 Adrian Chen, “Inside Facebook’s Outsourced Anti-Porn and Gore Brigade, Where ‘Camel Toes’ are More Offensive Than ‘Crushed Heads.’ Gawker.com, (2012), <http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads>.

2 The Guardian, “The Facebook Files,” The Guardian, <https://www.theguardian.com/news/series/facebook-files>

3 Julia Angwin, “Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children,” ProPublica, (2017), <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>

4 Channel 4, “Inside Facebook: Secrets of the Social Network, Channel 4 Dispatches,” Channel4.com (2018), <https://www.channel4.com/programmes/inside-facebook-secrets-of-a-social-network> [video]

Platforms are increasingly being called upon to make ethical decisions regarding speech. Within the United States, technology companies are largely given the leeway to set their own standards for content, as outlined in Section 230 of the Communications Decency Act. Section 230 provides platforms or “interactive computer services” with limited liability for most types of content posted by their users. However, there has been a rise in the public scrutiny of *how* platforms are making these decisions.⁵ This scrutiny has largely been directed toward Facebook, which has been the subject of numerous investigations following the rise of disinformation on the platform.⁶ However, attention has also been paid by media and government representatives to other large platforms such as YouTube and Twitter for both their role in facilitating disinformation and the power they have to filter online speech.⁷ Due to public attention concerning the use of platforms to organize and coordinate major politically extremist events like the Unite the Right rally in Charlottesville, Virginia, even smaller platforms, such as Patreon and Discord, have been called upon to draw clearer lines regarding what speech is allowed on their platform.⁸ Such lines are embedded in the rules outlined within the community guidelines documents that serve as the public-facing rules for individual users. While publicly accessible, this component of a site’s Terms of Service has been criticized for being too vague and opaque for platforms that govern speech for users all over the world.⁹ Less is known about how these policies are then deployed through a set of practices referred to as content moderation, which can include banning or removal of content or accounts, demonetization (on platforms like YouTube), de-ranking, or the inclusion of tags or warnings against problematic content.

Most platform companies keep their content moderation policies partially, if not mostly, hidden. Sarah T. Roberts, notes this contributes to a “logic of opacity” around social media moderation, that serves to make platforms *appear* objective, driven instead by “machine/machine-like rote behavior that removes any subjectivity and room for nuance,” inoculating companies from any “large-scale questioning of the policies and values governing

For all of their power to shape what we see and do not see online, the public is made most aware of the realities of **content moderation** only when such processes break down and reveal the complex and fraught decisions being made behind the scene.

⁵ 47 U.S. Code § 230

⁶ See The Guardian’s Facebook Files as an example.

⁷ House of Representatives Judiciary Committee, “Filtering Practices of Social Media Platforms,” Judiciary.House.Gov, (2018), <https://judiciary.house.gov/hearing/full-committee-hearing-filtering-practices-of-social-media-platforms/>.

⁸ Blake Montgomery, “PayPal, GoFundMe, and Patreon Banned a Bunch of People Associated With the Alt-Right. Here’s Why,” BuzzFeed News, (2017), <https://www.buzzfeednews.com/article/blakemontgomery/the-alt-right-has-a-payment-processor-problem>.

⁹ Gennie Gebhart, “Who Has Your Back? Censorship Edition 2018,” Electronic Frontier Foundation, (2018), <https://www.eff.org/who-has-your-back-2018>.

decision-making.”¹⁰ Scholars Kate Klonick¹¹ and Tarleton Gillespie¹² have also written about the invisible network of decision-making and platform governance that shapes what users see online. Klonick refers to platforms and moderation teams as the “New Governors” and argues they are part of “new triadic model of speech that sits between the state and speakers-publishers,” that she argues are “private self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free speech expectations of their users.”¹³ In his 2018 book, Gillespie describes the careful balancing acts of platforms as they weigh competing cultural values against each other to draw clear lines in blurry cultural environments.¹⁴ It has become clear that, for all of their power to shape what we see and do not see online, the public is made most aware of the realities of content moderation only when such processes *break down* and reveal the complex and fraught decisions being made behind the scene. This work has been undeniably helpful in shedding light on part of an industry that has been largely out-of-view.

This report builds on these understandings of content governance by looking at organizational dynamics within platform companies to examine how companies resolve creating consistent rules, while being contextual and localized, at different scales of operation. This report presents findings from interviews with 30 key stakeholders across platform companies, civil society, news media, fact-checking and verification organizations, and government. For the purposes of this report, I will primarily focus on responses from representatives of platform companies, both large and small, from a variety of positions on policy and technology teams.¹⁵ Interviews focused on how key policy personnel and product managers describe their approach to creating and enforcing standards for their communities. In addition, this report analyzes public statements made by platform representatives at the two *Content Moderation and Removal at Scale* conferences, hosted by Santa Clara University School of Law on February 2, 2018, and in Washington, D.C., on May 7, 2018. There are, of course, limitations with relying on statements made by public representatives of corporations; for instance, due to the opacity of much of content moderation policy, I often had to take their statements at face value. My interest, however, was more with how they *framed* the challenges they were facing publicly, which may contravene current dominant narratives in the public-at-large.

10 Sarah T. Roberts. (2018). “Digital detritus: ‘Error’ and the logic of opacity in social media content moderation.” *First Monday*, 23 (3-5). <https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649>

11 Kate Klonick, “The New Governors of Speech: The People, Rules, and Processes Governing Online Speech,” *Harvard Law Review*, 131 (2017): 1598-1620.

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

13 Kate Klonick, “The New Governors of Speech: The People, Rules, and Processes Governing Online Speech,” *Harvard Law Review*, 131 (2017): 1598-1620.

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

15 A full list of platforms we interviewed include Microsoft (Bing), Facebook, Google News, Reddit, Twitter, Vimeo, Patreon, Discord, Wikimedia, and Medium.

My focus is how key individuals in policy and technology development are creating standards for speech on their platform. I do not address how content is reported by users, nor labor concerns (which have been outlined by scholars such as Sarah T. Roberts¹⁶ and journalists such as Adrien Chen).¹⁷ Rather, I analyze how policy developments unfold within the organizational dynamics of specific platform companies that have emerged in prevalent discourse within content moderation communities. Based on this research, I separate platform companies into three distinct groups: (1) **Artisanal approaches** (a term which originated with the industry itself) such as Medium, Vimeo, Patreon, and Discord; (2) **Community-reliant approaches** such as Wikimedia and Reddit; and (3) **Industrial approaches**, which have been described by Tarleton Gillespie, and include major companies such as Facebook and YouTube. Across these categories, policy representatives navigate similar tensions, particularly between establishing *consistent rules for content and being sensitive to localized contexts for speech*. Taking a comparative approach across platforms, I was able to see when different models of content moderation policy tended to prioritize consistent rules or to be sensitive to the context of speech differently; for instance, industrial models prioritize consistency and artisanal models prioritize context. This tension between consistency and context reveals an important risk: What is lost when platforms are asked to formalize their rules *too quickly*? In many cases, the decisions made on content moderation standards are hard-coded into organizational practices, used to train thousands of new workers, and eventually transformed into automated flagging systems. At the same time, policies that rely too heavily on context risk being construed as targeting certain groups and individuals, particularly when these policies cannot be implemented at scale.

Understanding these differences and the nuances of each organization is helpful for determining both the expectations we should be placing on companies and the range of solutions that need to be brought to bear. This is important for the artisanal organizations that need to formalize their logic to address concerns more consistently. And it is also important for the industrial-sized operations that need to translate values into training and evaluations while being sensitive to the individual differences of content, like hate speech and newsworthiness. This white paper provides a cross-platform analysis of content moderation policy teams to explore how differences, in terms of size, value, and missions, inform approaches to content moderation. I will first unpack *why* different stakeholders (both internal to companies, and external, such as governments) are finding it necessary to differentiate between platform companies. Following that, I will explore the organizational dimensions of platform companies and the tensions artisanal, community-reliant, and industrial face as they work to establish consistent rules on their

¹⁶ Sarah T. Roberts. (2016). Commercial content moderation: Digital laborers' dirty work. In Noble, S.U. and Tynes, B. (Eds.), *The intersectional internet: Race, sex, class and culture online* (pp. 147-159). New York: Peter Lang. <https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1012&context=commpub>

¹⁷ Adrien Chen, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed," *Wired Magazine*, (2014), <https://www.wired.com/2014/10/content-moderation/>.

platforms while working to be sensitive to local, contextual concerns. Lastly, I will place organizational dynamics in conversation with existing and emerging regulation, particularly Section 230 within the United States and the NetzDG rule in Germany.

WHAT ARE PLATFORMS, ACCORDING TO WHOM?

The term “platform” is the most popular way to refer to online intermediaries, though it remains vague and unhelpful to governance or accountability conversations. This is because so many different companies working as online intermediaries, operating in so many different industries (from media, such as Facebook, to transportation, such as Uber) have, unhelpfully, adopted the term to appear neutral,¹⁸ evade regulatory classification,¹⁹ or avoid the normative or professional standards that may come with a given domain.²⁰ Over the past few years, several scholars have attempted to define and classify “platforms” in order to clarify questions of their regulation and oversight, though the term remains slippery.²¹ In compiling this report, I was guided by Tarleton Gillespie’s feature-based approach to defining platforms. Platforms are online sites and services that “host, organize, and circulate users’ shared content or social interactions for them,” without producing much of that content, built on an infrastructure for processing data (for multiple purposes), and that (most importantly) “moderate the content and activity of users.”²² Within these broad parameters, however, platforms still vary widely, both in terms of functions for users (e.g., search and social media) and industry (e.g., content and media, ride-sharing, payment apps, hostelry), making them difficult to define, regulate, and oversee.

Regulators should know that platform companies themselves are becoming wary of the slipperiness of the term. As behemoth companies like Facebook are continually excoriated for enabling the spread of false information, disinformation, and hate speech, other companies are seeking to distance themselves from what they feel are problems specific to social media. The companies I spoke with were incredibly diverse in terms of missions, business models, and size of both user bases and workers. Representatives frequently pointed to these diverse factors²³ to note that regulations that do not consider differences *between* platforms threaten to “lump all the technology together in ways that do not make

¹⁸ Tarleton Gillespie, “The Politics of Platforms,” *New Media & Society*, 12(3)(2010): 347-364.

¹⁹ Julia Cohen, “The Regulatory State in the Information Age,” *Theoretical Inquiries L. 17* (2016): 369.

²⁰ See Philip M. Napoli and Robyn Caplan, “When Media Companies Insist They Are Not Media Companies, Why They are Wrong, and Why That Matters,” *First Monday*, 22(5)(2017).

²¹ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018); Julia Cohen, “The Regulatory State in the Information Age,” *Theoretical Inquiries L. 17* (2016): 369.

²² Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018).

²³ A number of interviews with representatives touched on this. One representative from Vimeo noted that there would have to be a “different solution for every platform, because every platform is a little bit different....answering these really tricky, amorphous problems with really seeping overly broad legislation never worked out well, and I don’t see that changing in the future.”

good sense ... and fail to recognize that users will have very different purposes for accessing information on different types of platforms.”²⁴ The argument that legislation could be overly broad and unintentionally limit an industry is a familiar complaint from private companies worried about regulation. However, this concern was echoed by representatives from civil society organizations, like the Anti-Defamation League (ADL), who have sought to make platforms more accountable for issues like hate speech. Brittan Heller, the director of the Technology and Society project at ADL, acknowledged that each tech company faces unique problems “based on their business model, their target market, and the age, size, and maturity of their business.”²⁵ Platforms themselves worked to differentiate between their companies, tactically focusing primarily on **features of the technology, business model, and size of company** (particularly for their content moderation policy and enforcement teams). These attempts at differentiation highlight the difficulties in comparing platforms of different size and scope, creating new concerns for regulation that was previously constrained by the *lack of differentiation* contained within the term “platform.”²⁶ Attempts to highlight differences between their company, and others, might therefore be tactical but are important to keep in mind as regulators and civil society actors (and technology companies) work to draw boundaries around this industry.

Some representatives differentiated their platforms according to **the features of the technology** itself. A representative from the search engine Bing saw the function of search engines as being fundamentally different from social media and inherently more intertwined with democratic ideals, such as an *informed citizenry*.²⁷ He noted that “search is unique” because people *come to the site* with a query and are not pushed content as they are with social media. Furthermore, search engines do not host content on their own, but rather serve as a way users find and access third-party content. This distinction is important, because although Bing uses content moderation for features like autocorrect, according to Gillespie’s definition of platform, which “host, organize, and circulate,” search engines do not qualify as platforms in the same sense as a social media company, or even an online encyclopedia, such as Wikipedia.²⁸ This representative also differentiated search engines by noting the democratic potential for search because it is a “way that people can find divergent ways of thinking or unpopular points of view.” An advertiser I spoke with confirmed that advertisers themselves look at the features of search and social media *very differently* when deciding where to spend money; search engines were referred to as “intent platforms” where you can target advertising directly based on something you know an individual is looking for, versus social media, which is considered an “interruption platform”

²⁴ Interview with Michael Golebiewski, senior program manager at Microsoft Bing.

²⁵ Interview with Brittan Heller from the Anti-Defamation League.

²⁶ Julia Cohen, “The Regulatory State in the Information Age,” *Theoretical Inquiries L.* 17 (2016): 369.

²⁷ Interview with Michael Golebiewski, senior program manager at Microsoft Bing.

²⁸ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018).

where users have to be interrupted and persuaded to “doing something” different.²⁹ These experts made a connection echoed by scholar Philip M. Napoli in his (forthcoming) text: that social media has become a “push medium,” not unlike broadcast television, rather than the “pull medium” that was associated with the early internet.³⁰

Some companies differentiated themselves according to their **business model and revenue sources**. Representatives from companies that have eschewed advertising claimed that it is a reliance on advertising revenue and click-based metrics that makes platforms susceptible to manipulation. A representative from blogging platform Medium stressed that it is not advertising-based, but rather subscription-based, exempting it from the same risks and harms as platforms like Facebook and Google.³¹ Medium moved away from the ad-driven model intentionally; their representative explained, “There is a misalignment of incentives between what value the reader gets and what we’re getting as a platform, as a distributor trying to optimize for quality...it wasn’t going to work.”³² A representative from Wikimedia, a nonprofit with 501(c)(3) status in the United States, noted that its revenue model may insulate it from some of the problems that come from ad-revenue (though they were quick to acknowledge some minor issues with bots, and larger concerns with harassment). A representative from Vimeo echoed a similar sentiment, saying that it is ads that drive many of the content concerns around inflammatory content. “It really does change the game not being advertising supported, and not having that kind of direct revenue sharing [like YouTube]. We don’t have Logan Paul on our platform. We don’t have the same people who are seeking fame, and the advertising dollars attached to those high numbers.”³³ Instead, Vimeo also works as a subscription-based model, which they think attracts a different clientele — “A lot more professional users.” Though the platform has acknowledged issues with extremist content in the past,³⁴ they say they have largely been spared disinformation and “fake news” problems because of their financial model.

For the most part, however, the major point of differentiation was **scale of company**, in terms of user base, number of employees, and specifically the size of content moderation teams. Most of the work that has been done thus far by scholars³⁵ and journalists on content moderation for online content platforms have focused on three relatively large

²⁹ Interview with David Herrmann, founder of Social Outlier.

³⁰ Philip M. Napoli, (2019), *Media Technocracy: Algorithmic News, the Public Interest, and the Future of the Marketplace of Ideas*. (New York: Columbia Press, forthcoming).

³¹ Interview with Alex Feerst, head of legal at Medium.

³² Interview with Alex Feerst, head of legal for Medium.com

³³ Interview with Sean McGilvray, director of legal affairs and Trust and Safety at Vimeo.

³⁴ Natasha Lomas, “UK outs extremism blocking tool and could force tech firms to use it,” TechCrunch, (2018, <https://techcrunch.com/2018/02/13/uk-outs-extremism-blocking-tool-and-could-force-tech-firms-to-use-it/>).

³⁵ See Kate Klonick, “The New Governors of Speech: The People, Rules, and Processes Governing Online Speech,” *Harvard Law Review*, 131: 1598-1620. See also Tarleton Gillespie. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

US-based companies: Google (including YouTube), Twitter,³⁶ and Facebook. These companies operate on a massive scale within the United States and beyond. Google and Facebook have reported having billions of users.³⁷ Relatively, Twitter is much smaller – at 330 million active users – but is still viewed as important geopolitically, probably because of its role in political actors communicating with followers. Other platforms, like Discord or Reddit, are much smaller, though some also boast hundreds of millions of active monthly users.

In discussing what affected their approach to addressing content issues such as disinformation and hate speech, representatives of these platforms put the most emphasis on the size of the content moderation teams. Facebook has committed to having 20,000 workers in their content moderation and policy teams by the end of 2018,³⁸ and a representative for Google stated publicly that Google has 10,000 individuals working in content moderation for YouTube alone.³⁹ The team sizes for the other platforms I interviewed are much smaller. Patreon, a crowd-funding platform often used by alternative media producers, has a policy team of six full-time members, serving around 100,000 creators around the world.⁴⁰ Discord now has 10 full-time staff members who handle Trust and Safety and content moderation and removal, addressing between 600 and 800 videos per day.⁴¹ Representatives from Medium informed us their team size is anywhere from 5 to 7 individuals.⁴² Taking a much different approach, Wikimedia has a similar number of people developing policy in-house, but it has more than 100,000 volunteer editors who actively moderate for the site.⁴³ Reddit and Vimeo declined to give firm numbers but stressed that content moderation is only a fraction of their overall companies,⁴⁴ with Vimeo stating “we have a lean and mean team.”⁴⁵ This resource gap between team sizes and user bases are at the root of platform companies’ concerns over content moderation regulation. Evan Engstrom, executive director of Engine, an advocacy organization for startups, told me that laws will have to take into account these resource gaps, noting that even larger-scale platforms with more employees and better automated detection technologies vary widely between them: “Even

36 Twitter is actually a much smaller company in terms of user base (around 336 million monthly active users in the first quarter of 2018, according to Statista); however it has been included in things such as congressional hearings, most likely due to its use by key political figures (including President Donald Trump) and to concerns about Russian propaganda and bots on the platform. (See <https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/> for user statistics.)

37 According to Statista.com, Facebook had 2.23 billion active monthly users in Q2 of 2018 (<https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>), and Google’s social media video platform, YouTube, is expected to grow to 1.86 billion users in 2021 (<https://www.statista.com/statistics/805656/number-youtube-viewers-worldwide/>).

38 Anita Balakrishnan, “Facebook pledges to double its 10,000-person safety and security staff by end of 2018,” CNBC, (2017), <https://www.cnn.com/2017/10/31/facebook-senate-testimony-doubling-security-group-to-20000-in-2018.html>.

39 Comments by Nora Puckett at the Content Moderation at Scale conference in Washington, D.C., on May 7, 2018.

40 Interview with Colin Sullivan and Agnes Evrard from Patreon. Sullivan also expressed concern about a lack of resources to do this kind of work, saying, “The reason why I think size is a useful thing to think about is it’s a reflection of the resources available to that platform to actually comply with something.”

41 Email correspondence with Sean Li from Discord.

42 Interview with Alex Feerst, head of Legal from Medium.

43 Interview with Jacob Rogers, senior legal counsel for Wikimedia.

44 Interview with Jessica Ashoosh, director of Policy at Reddit Inc.

45 Interview with Sean McGilvray, director of legal affairs and Trust and Safety at Vimeo.

when you're talking about volume, high volume, within a specific industry where other companies might have a similar need. Not every company is going to have the resources to engage in these types of practices.”⁴⁶

In many cases, platforms are trying to achieve this balance and draw lines around “content” by first drawing lines around their user base and audience. In this sense, platforms are continuing their move from a “public square” model, where all speech (no matter how objectionable) is theoretically encouraged, toward one which embraces limited restrictions on its users. To some degree this has always been the case. All platforms had content rules at their outset (for instance, against illegal content and copyright protections which were introduced early on), and many platforms have been slowly incorporating new rules over the past two decades in response to growing public concerns about issues like terrorism and extremist content, revenge porn and harassment, or cyberbullying.⁴⁷ Hate speech, and now disinformation, are included within this list of concerns. According to one representative I spoke with, the old “public square” strategy has been criticized due to its lack of protection for marginalized individuals driven off sites against content like hate speech and conduct like harassment. They preferred to embrace a “curated community approach” which frames standards as “we are a group, and we have x, y, and z ethical codes of how we treat each other.” This sentiment was confirmed by Monika Bickert from Facebook who told us the company is not necessarily working to just “balance between safety and free speech,” but rather establish standards for speech, “to create a safe community.”⁴⁸

In making decisions about standards, platform companies are finding themselves caught between several competing tensions. On one hand, they are being asked to make their detailed content moderation rules available to the public.⁴⁹ On the other, they are being told that public rules can be easily gamed, with offenders carefully calibrating harassment to fall just short of moderation — a common problem, according to Brittan Heller of the ADL.⁵⁰ They are also increasingly having to make decisions in response to removal requests from foreign governments. This leaves them with a difficult choice, which is often framed by companies as a careful balance between respecting the sovereignty of a foreign nation or acceding to government censorship. The other option, to intervene and make decisions that preserve values they may hold (such as protecting the speech of LGBTQ users in areas where this speech is prohibited), was noted by one representative as being potentially interpreted as a Western corporation projecting their ideology abroad.⁵¹ However, across

⁴⁶ Interview with Evan Engstrom from Engine.

⁴⁷ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018).

⁴⁸ Interview with Monika Bickert, head of global policy management from Facebook.

⁴⁹ Russell Brandom, “New Rules Challenge Google and Facebook to Change the Way they Moderate Users,” *The Verge*, (2018), <https://www.theverge.com/2018/5/7/17328764/santa-clara-principles-platform-moderation-ban-google-facebook-twitter>.

⁵⁰ Interview with Brittan Heller from the Anti-Defamation League.

⁵¹ Interview with Alex Feerst, head of legal at Medium.

nearly all of the interviews with platform representatives, the greatest tension identified was how to be consistent in developing rules geared toward hate speech and disinformation, which tends to be highly contingent on local contexts and power dynamics.

Balancing Context Moderation with Consistent Moderation

The need to balance between consistency of rules, with being sensitive to local contexts, particularly for issues like hate speech and disinformation, is of particular concern when considering both the design of platforms and issues of scale.⁵² Preserving context is a major concern on platforms that tend to collapse it at every turn,⁵³ both in terms of how individuals receive information (a post from your friend and a message from a news agency tend to look fairly similar) across cultures with different histories and power dynamics as well as in the reception of information by both other users and moderators. At the same time, maintaining some level of consistency between decisions is necessary both philosophically, for instance in ensuring rules are not applied arbitrarily or ensuring some sense of “justice,” and practically, as thousands of workers are on-boarded to address content concerns. Maintaining this balance is not necessarily unique to platform companies; both national and international law has historically struggled with tailoring prohibitions against content, like hate speech, to varying traditions and histories, while remaining “law-like” and establishing minimum standards.⁵⁴ And yet, content policies are not law; they’re policy. This gives platform companies more leeway in applying their rules, while also largely hiding this process of policy development and enforcement from public view.

The need to balance between consistency of rules, with being sensitive to local contexts, particularly for issues like hate speech and disinformation, is of particular concern when considering both the design of platforms and issues of scale.

In content moderation, this challenge goes beyond developing rules to applying them. Particularly as a company scales, the review of content often happens *far outside the context* where it is produced. For a moderator to accurately assess whether content is hateful, they need to know the context of the content as it was made, including information about the individual making it, the target, and the environment, as well as linguistic or cultural clues they may not have access to (such as sarcasm, or newsworthiness). Moderators must also

⁵² Michael Herz and Peter Molnar, *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. (Cambridge University Press: Cambridge, UK, 2012).

⁵³ Henry Jenkins, *Convergence Culture: Where Old and New Media Collide*. (New York: New York University Press, 2006)., <https://www.hse.ru/data/2016/03/15/1127638366/Henry%20Jenkins%20Convergence%20culture%20where%20old%20and%20new%20media%20collide%20%202006.pdf>.

⁵⁴ Michael Herz and Peter Molnar, *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. (Cambridge University Press: Cambridge, UK, 2012).

be incredibly self-reflexive about *their own context*. As they work their way through the queue, moderators must not presume that viewers are seeing a particular post juxtaposed against the same hate speech, pornography, and crude humor they most recently reviewed.

Because of volume and job demands, moderators are often acting on all these factors in a few seconds (or less). One early employee of Facebook told us that making content moderation localized and responsive often required making decisions without sufficient information. He noted, “Who is historically disadvantaged with respect to whom is context dependent and situational,” citing an example of the difficulties of assessing hate speech against someone who is Japanese. Moderators must consider: “Are you historically advantaged because of Japanese imperialism in China, or are you historically disadvantaged because of the treatment of Japanese Americans in the United States?”⁵⁵ To address these context concerns at the scale at which Facebook was operating at the time (a tiny 70 million users compared to today’s 2 billion), “you would have to hire everybody in India to look at all the content that was uploaded, and you still wouldn’t be able to do it.” This has led a number of platforms – Twitter, for example – to look for other signals to moderate content (e.g., comments and interactions) that can be used to call attention to problematic behaviors. However, these too can create context concerns. One Twitter representative noted that likes and replies can mean many different things in different circumstances, making decisions difficult to automate:⁵⁶

People seek attention in similar ways, and a spammer seeking attention looks a lot like a rapper who’s trying to drop their latest mix tape, and the people that reply look a lot like someone who is trying to engage in a targeted harassment campaign. Just because something is coordinated, doesn’t mean it’s bad.

How a platform company balances these tensions depends significantly on the organization of the content moderation team. Between small, large, and medium-sized teams, we found several consistencies across how moderators were alerted to, and dealt with problematic content. There were, however, significant differences in how these teams were able to adapt to cultural variations, such as linguistic divides in content, as well as their capacity in using artificial intelligence to automate content flagging and removal. As we think about potential mechanisms to oversee these companies as they make important decisions about the future of speech online, we should pay attention to organizational dynamics and the tradeoffs companies are making, often hidden from public view.

⁵⁵ Interview with Craig Colgan (pseudonym), former employee at Facebook.

⁵⁶ Interview with Del Harvey, vice president of Trust & Safety at Twitter.

APPROACHES TO CONTENT AND CONTEXT MODERATION: Artisanal, Community-Reliant, and Industrial

When Craig Colgan⁵⁷ first joined Facebook in 2008, content problems like hate speech, propaganda, and “fake news” were not the main focus of his work.⁵⁸ Colgan initially took a job with the company in “user operations” — resetting passwords and helping users log in. The company was young, but it was growing quickly, expanding to new schools and markets. When Facebook lawyer Jud Hoffman started a new “elite” team called “site integrity operations,” Colgan saw an opportunity to get out of his tedious job.⁵⁹ This team was Facebook’s first effort to do content moderation, which, according to Colgan, was built to deal with an influx of “blown off heads and naked people and all kinds of unfortunate things.” Starting off as 12 people sitting in a room, the team worked together to deal with content that none of them had ever had to encounter in their offline lives. The team quickly grew, coming to absorb the support teams where Colgan had gotten his start. By the time he left, 12 people had become 500. By the end of 2018, this team will have expanded to 20,000 people, with offices all over the world, though it is likely this number reflects a large percentage of contract-based, outsourced workers, not contained within the company.⁶⁰

For platform companies, “content moderation” constitutes an incredibly diverse range of organizational structures, rules, and motivations.

⁵⁷ For select interviewees, we have used a pseudonym to protect their identities.

⁵⁸ Interview with Craig Colgan (pseudonym), former employee at Facebook.

⁵⁹ Kate Klonick, “The New Governors of Speech: The People, Rules, and Processes Governing Online Speech,” *Harvard Law Review*, 131 (2017): 1598-1620.

⁶⁰ Anita Balakrishnan, “Facebook pledges to double its 10,000-person safety and security staff by end of 2018,” *CNBC*, (2017), <https://www.cnbc.com/2017/10/31/facebook-senate-testimony-doubling-security-group-to-20000-in-2018.html>.

For platform companies, “content moderation” constitutes an incredibly diverse range of organizational structures, rules, and motivations. The teams that perform content moderation have a variety of names: Community and Fraud,⁶¹ Trust and Safety,⁶² and Law and Policy.⁶³ For larger companies, such as Facebook, these teams are made up of several smaller teams, like Product Policy, Community Operations, Community Integrity, and Escalations.⁶⁴ Smaller companies, which sometimes still serve millions of users, often have only one team for both policy development and policy enforcement. With teams as small as four, these organizations often rely on the same piecemeal or case-by-case approach to policy that Craig Colgan said was typical of Facebook in its early years.

Work like that done by Sarah T. Roberts, has explored, in particular, the labor and governance dynamics of what she refers to as “commercial content moderation,” describing a “set of practices with shared characteristics” where workers act as “digital gatekeepers for a platform...deciding what content will make it to the platform and what content will remain there.”⁶⁵ In contrast to Roberts’ work which focuses primarily on the major companies, this paper provides a comparative platform analysis across companies of different scales and governance models and explores differences between platform companies’ content moderation policies and practices.

We identify three major categories of platform companies according to their size, organization, and content moderation practices: (1) **The artisanal approach**, where case-by-case governance is normally performed by between 5 and 200 workers; (2) **Community-reliant approaches**, which typically combine formal policy made at the company level with volunteer moderators; and (3) **The industrial approach**, where tens of thousands of workers are employed to enforce rules made by a separate policy team. It is important to note these are *not new categories*, but rather emerge from the discourse used prevalently by platform representatives, and extends work done by Tarleton Gillespie in his book *Custodians of the Internet*, in which he notes the role that scale plays in approaches to moderation.⁶⁶

⁶¹ Comments made by Casey Burton, senior corporate counsel at Match, at the Content Moderation at Scale event in Washington, D.C., on May 7, 2018.

⁶² Comments made by Del Harvey at the Content Moderation at Scale event in Washington, D.C., May 7, 2018. Analysis taken from comments made by practitioners at the Content Moderation at Scale event in Washington, D.C.

⁶³ Analysis taken from comments made by practitioners at the Content Moderation at Scale event in Washington, D.C.

⁶⁴ Part of four parts that make up Facebook’s content moderation team, which includes Product Policy, Community Operations, Community Integrity, Escalations.

⁶⁵ Sarah T. Roberts. (2016). “Commercial Content Moderation: Digital Laborers’ Dirty Work.” n Noble, S.U. and Tynes, B. (Eds.), *The intersectional internet: Race, sex, class and culture online* (pp. 147-159). New York: Peter Lang. <https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1012&context=commpub>

⁶⁶ Gillespie’s full quote is: “This is not just a difference of size, it is fundamentally a different problem. For large-scale platforms, moderation is industrial, not artisanal.”

The Artisanal Approach

The field of artisanal organizations operating in the area of content is incredibly diverse and includes large-scale chat platforms like Discord, smaller user-generated content sites like Medium or Vimeo, and payment apps servicing media and content producers like Patreon. Therefore, while not as prominent as Facebook or Google, collectively these services are some of the main access points to information for many of the world's internet users. And while the vast majority of content moderation teams use the artisanal approach, speaking with representatives from Medium, Discord, Patreon, and Vimeo, it was apparent much academic and policy work on content moderation does not address those platforms whose operations are artisanal.

Smaller-scale operations most often emphasize a hands-on approach to content moderation. Alex Feerst, head of legal for the blogging platform Medium, referred to their approach as “artisanal,” or (being tongue-in-cheek) as “small-batch,” to note that despite their more than 80 million users, their moderation approach is still done manually, “by human beings.” This is a stark contrast to the way a moderation company such as Facebook has been described in media reports.⁶⁷ Other artisanal companies used similar language; Sean McGilvray, director of Legal Affairs and Trust and Safety at Vimeo, noted it reviews all flagged content “in-house” and further distinguished their practices from a company like Facebook: “We don’t have any third-party contractors. We don’t have a warehouse of people. We’re viewing every reported video.”⁶⁸ Artisanal content moderation teams are therefore not just distinguished merely by their smaller size, but the degree to which content moderation is done *in-house* by employees within the company (opposed to contractors), and the limited use of automation and access algorithmic detection technologies.

Artisanal approaches are also limited in their use of automation. Representatives note how such technologies are used to help surface content for human review but not to make moderation decisions. Feerst stressed that outside of these limited categories, at Medium, “a human being looks at everything.”⁶⁹ Vimeo, known for hosting more artistic content, said that automated tools present problems when prohibited content types are less clear-cut. This platform prohibits pornography but allows for nudity for artistic purposes. He said automated tools, even ones incorporating artificial intelligence and machine learning, cannot navigate the complexities involved in making this sort of decision, stating, “You’re never going to train an AI to recognize that. You can barely train a human to do it consistently, so I don’t see a machine taking over that initiative any time soon.”

⁶⁷ For an example, see Slate’s coverage, calling Facebook and Google’s content moderators an “army.” <https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html>.

⁶⁸ Interview with Sean McGilvray, director of Legal Affairs and Trust and Safety at Vimeo.

⁶⁹ It should be noted that the major companies are also stressing that they use limited automation, but as reports such as the Facebook transparency report shows, they are increasingly using automated detection technologies for purposes outside of spam, child pornography, or even terrorist accounts, extending its use to violence, hate speech, and fake accounts.

Companies using the artisanal approach often pride themselves for being able to be more responsive to the context in which speech was made.

Though these companies, like the larger ones, are wary of providing exact numbers, they emphasized just *how small* their teams are, both in developing policy and enforcing it. Content moderation teams can range from as incredibly small as four people, up to a few hundred. The smaller the team, the more “small-batch” the decision-making, with formalized processes becoming more concrete and ingrained as they grow larger. The Medium Trust and Safety team consists of between 5 and 10 people at any given time, which includes Feerst, head of legal, as well as engineers who design tools to scale up flagging and reviewing. Companies like Patreon and Discord have similarly small teams. Patreon has a policy team of four full-time members, serving around 50,000 creators around the world. Discord also has four full-time staff, addressing around 300 reports per day. In these settings, policy development is often occurring in the same space as policy enforcement, with the same few people responsible for responding to content and developing the rules for doing so. This means that content types that defy categories can often be discussed in-depth, using tactics like debates, or mock trials to better understand difficult-to-classify content.⁷⁰ One legal counsel compared the model they took to a “common-law system” based on precedent, while others described a process similar to a *grounded theory* approach,⁷¹ a methodology used in the social sciences to inductively build up categories, through the aggregation of individual cases or data points.⁷²

Companies using the artisanal approach often pride themselves for being able to be more responsive to the context in which speech was made. Though some of these companies may have millions of users, many artisanal organizations claim to have a lower rate of reports for problematic content. Representatives attribute this to a combination of factors: business models, user populations, and content type. The lower reporting rate means these platform companies can take more time to review each post. Team members at Discord often spend “10 to 20 minutes” looking at the “servers” (the subcommunities in the Discord universe) where violations are occurring as opposed to the seconds given to a moderator at Facebook or YouTube.⁷³ Though this approach gives moderators more time to determine context and react accordingly, other resource issues, such as limited language capacity outside of English, still constrain the ability of these companies to fully understand cultural and political contexts. They often have to be creative in how they

⁷⁰ Interview with Alex Feerst, head of Legal at Medium.

⁷¹ Barney Glaser, and Anselm Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*, (New York: Aldine de Gruyter, 1967).

⁷² Interview with Craig Colgan (pseudonym), former employee of Facebook.

⁷³ Interview with Sean Li, director of Trust and Safety at Discord.

localize their responses to issues that require localized knowledge. Some companies use translation tools (like Google Translate), while others contract specific human-led translation services.⁷⁴ One company noted they make use of an informal network of experts – mostly academics – to solicit feedback on potential decisions.

The experiences of artisanal content moderation teams were consistently similar to those of early policy employees at major online content platforms. Craig Colgan described his early experiences in Facebook’s content moderation division as being similar to that of smaller platforms now, with a one-page list of informal rules, and workers “sitting in the same room or two rooms.”⁷⁵ After a while, however, it became clear this model “wasn’t going to scale,” so Facebook worked to develop a more “comprehensive and systemic set of standards.” Nicole Wong, former vice president and deputy legal counsel at Google offered a similar account of Google’s earliest days in 2004, on their limited content platforms: “There was effectively no moderation. There were customer support people who answered questions, but there were no true moderation policies, per se, other than for copyright and child pornography.”⁷⁶ Over time, she said, key events (such as the Yahoo v. LICRA case, regarding the availability of Nazi paraphernalia globally) played an important role in developing policies, with formalization occurring over time as the platform grew.

Though they had fewer resources, they also had fewer reports, and arguably, lower stakes, reputationally and financially, if they failed to make a good decision. These companies also seemed to place an emphasis on learning from each case, developing rules more slowly over time, with little worry they would need to construct a black-and-white rule to be deployed by an algorithm (largely because they lack the requisite financial and technical resources, including enough data to train an algorithmic model). Because of this, their rules tend to be opaque and less consistent, leading to concerns about transparency and fairness in their application. Additionally, from our interviews, representatives also noted the significant organizational costs, such as employee resources to have debates or deep discussions, which must happen when taking on each case on its own, without having a set of formal overarching rules used by larger companies such as Facebook and Google to guide individual decisions. As these companies attempt to scale, and if they become subject to a rule like the NetzDG in Germany, they will have to actively add employees and formalize rules much faster than what was afforded existing major companies such as Facebook and Google.

⁷⁴ Interview with Sean Li, director of Trust and Safety at Discord.

⁷⁵ Interview with Craig Colgan (pseudonym), former employee of Facebook.

⁷⁶ Interview with Nicole Wong, former vice president and deputy general counsel at Google.

The Community-Reliant Approach

Community-reliant organizations are platform companies that have created structures for large groups of volunteer users to implement and add to the overarching policy decisions of a small team employed by the company. Because the users are doing a significant portion of the actual moderation, these organizations cannot be neatly understood by team size. Instead, these platforms must address a special set of moderation concerns that come from their entanglement with their volunteer moderators. Perhaps the most significant amount of existing scholarly research on moderation concerns such platforms — those with a combined approach between a parent company (in the case of Reddit) or organization (such as the nonprofit Wikimedia) and a large body of users. Scholars like Paul B. de Laat⁷⁷ and Stuart Geiger⁷⁸ have written extensively on how members of the Wikipedia community moderate both content and the conduct of its members (especially in curtailing harassment, a major issue for the Wikimedia platform). In his piece “The Virtues of Moderation,” James Grimmelman analyzes Wikipedia and Reddit as sites of “distributed moderation,” and focuses on the role of norm-setting between members of the sites’ subcommunities.⁷⁹

Though nearly every speech platform relies on its users to aid in the process of moderation – primarily by flagging content for review – platforms like Wikimedia (the parent organization of Wikipedia) and Reddit rely on volunteer moderators much more substantially. Both organizations separate powers between the parent organization and its subcommunities, with the parent organization setting overarching norms and standards, which can be added onto by subcommunities contained within the site. Wikimedia is unique in their approach as both a nonprofit and an open-community-based model. Though representatives from the organization said they are willing to create policies regarding the *conduct* of editors (such as harassment), *content* decisions (aside from those on illegal or copyrighted content) largely remain out of their control.⁸⁰ Content decisions are left almost entirely up to the discretion of admins and editors, a volunteer workforce of tens of thousands of active users who set rules “through a consensus process done over many years of building up policies,” according to one Wikimedia representative.⁸¹ When someone isn’t following a policy, or if there’s a conflict between editors, individuals can ask for review from admins or other third parties, but the Wikimedia Foundation is

⁷⁷ Paul B. de Laat, “Coercion or empowerment? Moderation of content in Wikipedia as ‘essentially contested’ bureaucratic rules,” *Ethics and Information Technology*, 14(2) (2012): 123-135, <https://link.springer.com/article/10.1007/s10676-012-9289-7>.

⁷⁸ For example: R. Stuart. Geiger, “Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space,” *Information, Communication, and Society* 19(6) (2012), <http://www.stuartgeiger.com/blockbot-ics.pdf>.

⁷⁹ James Grimmelman, “The Virtues of Moderation,” *The Yale Journal of Law & Technology* (17) (2015), <https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2620&context=facpub>. See also J. Nathan Matias, “What Just happened on Reddit? Understanding the Moderator Blackout,” *SocialMediaCollective.org*, (2015), <https://socialmediacollective.org/2015/07/09/what-just-happened-on-reddit-understanding-the-moderator-blackout/>.

⁸⁰ Interview with Jacob Rogers, senior legal counsel at Wikimedia Foundation.

⁸¹ Interview with Jacob Rogers, senior legal counsel at Wikimedia Foundation.

largely removed. A representative noted that the Wikimedia Foundation is only notified of violations if volunteers are having trouble addressing issues themselves, and “even then, we can’t always answer those questions either.”

A representative for Reddit compared their model to a “federal system” with baseline site-wide rules that must be obeyed by smaller subcommunities but can also be extended according to the discretion of sub-community moderators.⁸² This representative compared it to the distribution of federal and state powers within the United States: “Similarly to the U.S. Constitution, states are not allowed to have laws that are in contravention of the Constitution, and subreddits are not allowed to have rules that are in contravention of our site-wide rules.” These rules are high-level and general, prohibiting illegal content like sexually suggestive content involving minors, as well as involuntary pornography, harassment, the posting of personal or confidential information, and content that encourages or incites individuals to violence.⁸³ Like other platform companies, Reddit did not provide concrete numbers for their in-house content moderation team but said “around 10% of the company is dedicated to fighting abusive content on the site, whether that abusive content is bad content posted by users or spam or bots.”⁸⁴ The total company size is around 400. The number of volunteer moderators on the site is much larger than the portion of the company dedicated to moderation. In a study conducted for Microsoft Research in 2015, J. Nathan Matias found 91,563 unique moderator accounts, with an average of 5 moderators per subreddit.⁸⁵

Though the community-reliant model creates some problems as subcommunities with different norms interact on each site, policy managers from these companies/organizations tended to believe that enabling communities to make *their own rules* enabled a greater sensitivity to potential cultural context concerns that the parent organizations could not address on their own, which, it should be noted, also conveniently serves their business aims. Reddit in particular tends to only get involved if users disobey site-wide rules, leaving enforcement up to individual communities. Wikimedia’s policy to leave content decisions to the admins and editors of individual sites means that policies can often be quite localized, even though the foundation is based within San Francisco and complies only with laws that apply to them, such as the Digital Millennium Copyright Act. Jacob Rogers, senior legal counsel for the Wikimedia Foundation, made the case that the Wikimedia Foundation is not legally *required* to comply with non-United States laws, even if a majority of the Wikipedia sites users are coming from a specific region (Wikipedia sites are normally divided by language, not country). However, he notes that some regional

⁸² Interview with Jessica Ashoosh, director of Policy at Reddit.

⁸³ Reddit Content Policy. <https://www.redditinc.com/policies/content-policy>. Accessed July 20, 2018.

⁸⁴ Interview with Jessica Ashoosh, director of Policy at Reddit.

⁸⁵ J. Nathan Matias, “What Just Happened on Reddit? Understanding the Moderator Blackout,” SocialMediaCollective.org, (2015), <https://socialmediacollective.org/2015/07/09/what-just-happened-on-reddit-understanding-the-moderator-blackout/#howmany>.

or national laws are so entrenched that language-specific Wikipedia sites do tend to adopt them as norms regardless. As an example, he mentioned that the German language Wikipedia has policies about Nazi-related content, but Rogers clarified that there is no legal obligation to remove the content.

Though it affords greater leeway to individual communities to create and enforce their own standards, these models are typically criticized for relying primarily on volunteers, who are not compensated for the work they do.⁸⁶ J. Nathan Matias has noted that relying on this unpaid and volunteer labor upholds “platform funding models” through reducing labor costs, and in policy, can even “limit [the platforms’] regulatory liability for conduct on their service while positioning themselves as champions of free expression and cultural generativity.”⁸⁷ Users of these sites who volunteer to take on additional roles moderating or developing policy (or engaging in other moderation activities, like up-or-down voting or flagging) invest their own time and resources into the site, which can lead to complicated power dynamics when the parent organization makes an overarching change. In this sense as well, the relationship between volunteer workers and parent organization can be quite adversarial, with volunteers, who often have their own vision for the site, sometimes pushing back heavily against broad-level rules. This was seen in the case of interim CEO Ellen Pao with Reddit, who was the target of significant anger and harassment from Reddit users when decisions were made to ban harassing subreddits like “/r/fatpeoplehate.”⁸⁸ This revolt eventually led to Pao’s resignation from the leadership position, raising questions about whether existing power inequalities among users and moderators of community-reliant sites doom content policies designed to combat harassment against (and thus increase speech of) marginalized communities.⁸⁹

The most relevant feature of community-reliant organizations is not just the size of the policy development team, nor the number of enforcers – though relying on volunteers often means they can surpass major companies such as Facebook and Google in terms of number of moderators⁹⁰ – but rather the relationship between the parent organization and its volunteer moderators. In these organizations, a base set of rules is added on to by subcommunities, each with their own norms and standards. At least in a theoretical sense, the balance between making consistent rules and being sensitive to context is partly solved by this arrangement. The central organization sets minimum standards

⁸⁶ Sarah T. Roberts, “Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation,” Dissertation, University of Illinois at Urbana-Champaign (2014), https://www.ideals.illinois.edu/bitstream/handle/2142/50401/Sarah_Roberts.pdf?sequence=1.

⁸⁷ J. Nathan Matias, “The Civic Labor of Online Moderators,” *The Platform Society*, (2016), <https://pdfs.semanticscholar.org/0a11/fb93ada453ec27c7fe-c63e69508e7e6201cd.pdf>. For another entry into this discussion, see Lisa Margonelli’s (1999) piece, “Inside AOL’s ‘Cyber Sweatshop’” for *Wired*, which examined the US Department of Labor investigation into the unpaid labor of content moderation at the company.

⁸⁸ Alex Abad-Santos, “The Reddit Revolt That Led to Pao’s Resignation, Explained,” *Vox*, (2015), <https://www.vox.com/2015/7/8/8914661/reddit-victoria-protest>.

⁸⁹ Davey Alba, “Ellen Pao Steps Down as CEO After Reddit Revolt,” *Wired*, (2015), <https://www.wired.com/2015/07/reddit-ceo-ellen-pao-steps-down-huffman-replacement/>. See also J. Nathan Matias, “Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout.” CHI’16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, CA, May 7-12 (2016).

Interview with Jacob Rogers from Wikimedia Foundation.

they should be able to enforce with subcommunities responsible for adopting specific rules in relation to their own communities. In practice, however, this seems to be much more complicated. Controversy and conflict about those base rules remain and, as in the case of Ellen Pao, the amount of ownership felt by unpaid moderators working for the company, can create divisions between the central organization and its users, as well as between different subcommunities that are at odds. In the end, the central organization often acts as a key mediator between communities, placing them in the same difficult position of drawing lines and setting clear standards as other companies are in setting rules about content.

The Industrial Approach

Public interest in content moderation has typically focused on a small number of larger companies – mainly Facebook and Google (YouTube primarily) – that have been called “industrial”⁹¹ due to their scale and number of users, the size of their content moderation teams, their operationalizing of rules, and the separation between policy and enforcement at their companies. These companies tend to have more resources and are continuing to add employees in content moderation rapidly. In light of concerns about disinformation and propaganda, Facebook has committed to expanding its content moderation team (including contracted organizations) from 10,000 to 20,000 by the end of 2018.⁹² At both Facebook and Google, policy development is separated from policy enforcement.⁹³ Companies like YouTube (owned by Google) also separate teams in terms of expertise and language fluency so that content can be funneled to the appropriate individuals to moderate.⁹⁴

One of our respondents said the goal for these companies is to create a “decision factory,”

These larger companies typically began content moderation in the artisanal model and used this period of experimentation to develop rules that become more formalized, static, and inflexible. Part of this formalization has occurred due to rapid growth and a need to train workers who are being on-boarded *en masse*. These workers often make decisions about content far away from the context of the initial speech. Ensuring fair and consistent decisions often means breaking complex philosophical ideals about what constitutes

⁹¹ A number of our interviewees mentioned this word, but it was also used by Tarleton Gillespie in his book *Custodians of the Internet*.

⁹² Anita Balakrishnan, “Facebook pledges to double its 10,000-person safety and security staff by end of 2018,” CNBC, (2017), <https://www.cnbc.com/2017/10/31/facebook-senate-testimony-doubling-security-group-to-20000-in-2018.html>.

⁹³ Part of 4 parts that make up Facebook’s content moderation team, which includes Product Policy, Community Operations, Community Integrity, Escalations.

⁹⁴ Comments made by Nora Puckett at the Content Moderation at Scale conference on May 7th, 2018.

harassment, hate, or truth into small components that are more likely to be interpretable. One of our respondents said the goal for these companies is to create a “decision factory,” which resembles more a “Toyota factory than it does a courtroom, in terms of the actual moderation.”⁹⁵ Complex concepts like harassment or hate speech are operationalized to make the application of these rules more consistent across the company.⁹⁶ He noted the approach as “trying to take a complex thing, and break it into extremely small parts, so that you can routinize doing it over, and over, and over again.” In this sense, industrial organizations are large-scale bureaucracies, with highly specialized teams and distributions of responsibilities and powers. As Gillespie has noted, this spread of labor, often widely dispersed throughout the company and the globe, also leads to logistical challenges in transmitting information about changing policies. This includes information about policies’ effectiveness or accuracy being conveyed back to policymakers at the company.⁹⁷

These companies operationalize their content policies because of their size; the sheer scale of content that needs to be reviewed is hard to even fathom. According to Nora Puckett, the YouTube representative at the 2018 Content Moderation at Scale Conference in Washington, D.C., in the fourth quarter of 2017, YouTube removed 8.2 million videos from 28 million videos flagged, which included 6.5 million videos flagged by automated means, 1.1 million flagged by trusted users, and 400,000 flagged by regular users. According to that same representative, YouTube has 10,000 workers in their content moderation teams. Twitter, which is dwarfed by behemoths Facebook and Google, still has 330 million monthly users and billions of tweets per week. At the Content Moderation at Scale event, Del Harvey, vice president of Trust & Safety at Twitter, noted that with this kind of scale, catching 99.9% of bad content still means that tens of thousands of problematic tweets remain.⁹⁸

Industrial content moderation teams are increasingly using automated tools to flag content such as hate speech. Both Facebook and YouTube have disclosed that they are now using algorithms to find offensive content and take it down using “detection technology” before such content is even flagged by users, though this content is still subject to human review.⁹⁹ In Facebook’s recent disclosures of automated takedowns of content, they stated that they had reported high rates of success for this detection technology in the areas of graphic violence (86%), nudity and adult content (96%), and spam (100%). For hate speech, the rate of success of automated technology is lower, but still significant, with detection technologies finding and flagging “around 38% of the content they took action

⁹⁵ Interview with Craig Colgan (pseudonym), former employee of Facebook.

⁹⁶ Tarleton Gillespie discusses this same process in his book *Custodians of the Internet*.

⁹⁷ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018).

⁹⁸ Comments made by Del Harvey at the Content Moderation at Scale event in Washington, D.C.

⁹⁹ Facebook, “Hate Speech,” [Transparency.Facebook.com](https://transparency.facebook.com/community-standards-enforcement#hate-speech), (2018), <https://transparency.facebook.com/community-standards-enforcement#hate-speech> Retrieved July 31, 2018..

on for hate speech, through automated means.”¹⁰⁰ Therefore, for content types that are considered less ethically ambiguous by the company, like spam/malware, child pornography, and terrorist propaganda (which requires its own investigation as to how companies are categorizing this type of content and the extent of false positives), rates of automated takedown are significantly higher. As evidenced by Facebook’s detection of issues like hate speech, companies may be exploring the use of automated technologies in these other domains.

When global platforms grow to be the size of Facebook or YouTube, maintaining consistency in decision-making is often done at the expense of being localized or contextual. This can cause problems in the case of content like hate speech, discrimination, or disinformation when making a moderation decision depends on particular cultural and political environments. Perhaps because of this, platforms of this size tend to collapse contexts in favor of establishing global rules that make little sense when applied to content across vastly different cultural and political contexts around the world. This can, at times, have significant negative impact on marginalized groups. Julia Angwin criticized this type of policy practice when Facebook attempted to implement a policy that incorporated notions of intersectionality divorced from existing power arrangements, essentially protecting the hegemonic groups of *White* and *men*, but not “Black children.”¹⁰¹ Her work demonstrated that attempts at universal anti-discrimination rules too often do not account for power differences along racial and gender lines. In other instances, this can mean the Venus of Willendorf is accidentally censored for being too “pornographic.”¹⁰² Such failures to address context issues can also lead to serious consequences. This has been seen, tragically, in the violence that has ensued in Myanmar, which has been arguably fueled by disinformation and hate speech that spread over both the Facebook platform and its messaging application WhatsApp.¹⁰³ In April 2018, Facebook CEO Mark Zuckerberg acknowledged that the company lacked the linguistic and cultural resources to quell hate speech in the region.¹⁰⁴ Though he pledged to hire more Burmese speakers, Reuters has reported that hate speech directed against the Rohingya community remains rampant across Facebook-owned and -operated products.¹⁰⁵

¹⁰⁰ Facebook, “Hate Speech,” Transparency.Facebook.com, (2018), <https://transparency.facebook.com/community-standards-enforcement#hate-speech> Retrieved July 31, 2018.

¹⁰¹ Julia Angwin, “Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children,” ProPublica, (2017), <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

¹⁰² Aimee Dawson, “Facebook censors 30,000-year-old Venus of Willendorf as ‘pornographic,’” The Art Newspaper (2018), <https://www.theartnewspaper.com/news/facebook-censors-famous-30-000-year-old-nude-statue-as-pornographic>.

¹⁰³ Anthony Kuhn, “Activists in Myanmar Say Facebook Needs to do More to Quell Hate Speech,” NPR (2018), <https://www.npr.org/2018/06/14/619488792/activists-in-myanmar-say-facebook-needs-to-do-more-to-quell-hate-speech>.

¹⁰⁴ Steve Stecklow, “Why Facebook is Losing the War on Hate Speech in Myanmar,” Reuters (2018), <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

¹⁰⁵ Steve Stecklow, “Why Facebook is Losing the War on Hate Speech in Myanmar,” Reuters (2018), <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

CONSISTENT VERSUS CONTEXTUAL PLATFORM REGULATION

The question of resource gaps *between companies*, whether artisanal, community-reliant, or industrial, has become higher-stakes following the creation of new regulation in non-US countries. Germany recently passed a law titled the Netzwerkdurchsetzungsgesetz (NetzDG) that places the same restrictions on hate speech online that have been placed on other media in the past.¹⁰⁶ The law, passed in fall 2017, applies to “profit-making” social media sites and “platforms offering journalistic or editorial content” with over two million registered users that receive more than 100 complaints per calendar year about “unlawful content.”¹⁰⁷ The law places obligations for reporting on the handling of illegal content, transparent procedures for how complaints are addressed, as well as auditing guidelines that are directed primarily at the organization of content moderation and trust and safety teams. The law specifies that a platform must have two million users to be subject to these provisions, and carves out protections for sites under this threshold, as well as nonprofit-making entities (potentially incentivizing more nonprofit social media companies).¹⁰⁸ Engstrom noted that it is easy for companies or even individuals to pass that two million user threshold, saying, “Just look at the followers that Instagram celebrities have. Millions and millions, just one person.”¹⁰⁹ Of the companies we spoke with, a number were concerned about this part of the NetzDG, particularly because the law drew clear lines where none yet existed and implied the use of automation by organizations to catch illegal content, which also does not yet exist.

US law currently does not distinguish platforms according to size, but instead between the designation of “interactive computer services” and “publisher.”¹¹⁰ Platforms (or rather “interactive computer services”) retain immunity for *most* types of non-illegal, non-copyrighted material due to Section 230 of the Communications Decency Act. Further, due to a “Good Samaritan” provision within the law, they are allowed to voluntarily “restrict

¹⁰⁶ Claudia Haupt, “Online Speech Regulation: A Comparative Perspective,” Presented at the American Political Science Association, August (2018).

¹⁰⁷ Gesetz zur Verbesserung der Rechtsdurchsetzung in den sozialen Netzwerken [Act to Improve Enforcement of Law in the Social Networks], BGBl. I, S. 3352 of Sept. 1, 2017, (Netzwerkdurchsetzungsgesetz, “NetzDG”). English translation by the Federal Ministry of Justice available at https://www.bmjv.de/Shared-Docs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=92275CBB36905E837DBADFEEE79A0533.1_cid324?__blob=publicationFile&v=2

¹⁰⁸ It’s not yet specified as to whether this means two million German users, or users worldwide. This ambiguity was cited by many of our respondents as the reason why they would not be subject to the law, while others were confused as to whether it would apply.

¹⁰⁹ Interview with Evan Engstrom, executive director at Engine.

¹¹⁰ Philip Napoli and Robyn Caplan, “Platform or Publisher?” International Institute of Communications (2017), <http://www.iicom.org/intermedia/intermedia-past-issues/intermedia-jan-2017/platform-or-publisher>.

access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected.”¹¹¹ This law gives platforms the leeway to develop their own community guidelines and enforce them as they see fit. Proponents of the law have made the case that the tech industry as we know it would not exist without this provision. Eric Goldman has called it a “globally unique solution” which has given the United States a competitive advantage when it comes to the internet.¹¹² Legal scholar Jack Balkin wrote in 2014 that he considered this rule to be “among the most important protections of free expression in the United States in the digital age.”¹¹³ Critics of the law say that the liability shield for platforms is too broad,¹¹¹ and that new exceptions and regulations need to be added to reduce defamation online.¹¹² Other critics have noted that the “Good Samaritan” provision has merely placed the public burden of regulation of speech onto platforms, with few formal mechanisms for oversight and accountability.¹¹⁶

Within the United States, Section 230 of the Communications Decency Act provides platforms like those discussed above with the freedom to organize their content moderation teams as they see fit, as long as they are taking care to remove copyright protected and illegal content. As platforms deploy the other right given to them by Section 230 and the “Good Samaritan” provision, platforms told us they are finding it difficult to draw lines in ways that make sense both ethically and organizationally. Though policy representatives we interviewed spoke favorably of the limited liability provision of Section 230, it became apparent that approaches to speech are as much of an organizational concern for these companies as they are a regulatory concern. Many of these representatives cited a need for clear guidelines and procedures that could guide some of the messy decisions they

“It creates a situation of existential freedom,” noting that “when you are floating in a void with no writing, you’re like, ‘Well, we’re here, and we have no idea how to orient. Let’s figure out a moral compass.’”

¹¹¹ U.S.C. Section 230 (c)(2)(A).

¹¹² Comments made at the Content Moderation at Scale Conference in Washington, D.C., on May 7, 2018.

¹¹³ James Grimmelman, “The Virtues of Moderation,” Yale Journal of Law and Technology, 17 (2015), <http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt>.

¹¹⁴ Kathleen Ann Ruane, “How Broad a Shield? A Brief Overview of Section 230 of the Communications Decency Act,” Congressional Research Service, (2018), <https://fas.org/sgp/crs/misc/LSB10082.pdf>.

¹¹⁵ Ann Bartow, “Section 230 Keeps Platforms for Defamation and Threats Highly Profitable,” The Record (2017), <https://www.law.com/therecorder/sites/therecorder/2017/11/10/section-230-keeps-platforms-for-defamation-and-threats-highly-profitable/?slreturn=20180629124258>.

¹¹⁶ Julie Adler, “The Public’s Burden in a Digital Age; Intermediaries and the Privatization of Internet Censorship,” J.L. & Poly. 20(2011).

make about content like hate speech and disinformation. Many also cited a need for more clarity around Section 230, like there is for the Digital Millennium Copyright Act, which provides platforms with procedures for notice-and-takedown, mechanisms for violators to respond (though not required), with clear outcomes for repeat offenders.¹¹⁷

Company representatives we spoke with repeatedly told us that the lack of clear guidelines within Section 230 is quite limiting. One respondent said, “It creates a situation of existential freedom,” noting that “when you are floating in a void with no writing, you’re like, ‘Well, we’re here, and we have no idea how to orient. Let’s figure out a moral compass.’”¹¹⁸ He suspected that was why platform companies place such a large emphasis on their “mission,”¹¹⁹ because it’s an attempt to create a “central framework, a shared premise that you can use to resolve decisions within the company.”¹²⁰ Even those who maintained a clearer commitment to Section 230, such as Evan Engstrom from the startup advocacy group Engine, noted that though he is not in favor of a “burden the messenger approach” that would hold platforms liable, giving platforms “clear guidelines about what I need to do when I come across something that I want to get rid of” is critical.¹²¹ Legal counsel for one platform also noted Section 230’s major issue is that “it provides the backdrop that allows platforms to moderate content, but does not give platforms guidance on what kinds of rules to adopt or ways to implement them.” This means that companies alone must determine how to develop and enforce rules, and how they’ll apply this at scale, as the cultural environment shifts quickly. At the same time, asking these companies about issues they have experienced with the Digital Millennium Copyright Act (DMCA) demonstrates the cleavages between organizations with different resources; for instance, a company the size of Medium must address DMCA concerns manually, while Google and Facebook have the resources to automate takedowns. Regardless, despite fears of a slippery slope toward the end of Section 230 that were expressed by startup advocate Engstrom, US courts have consistently interpreted it to give platforms almost complete immunity, so it is unlikely that major changes to the law will occur.¹²²

According to legal counsel for one platform, outside of the US, “the law sometimes has more detailed requirements depending on what type of content is at issue.” The NetzDG law in Germany provides much more in terms of procedure for addressing complaints. It places much clearer obligations for reporting on the handling of illegal content, transparent procedures for how complaints are addressed, as well as auditing guidelines, that are directed primarily on the organization of content moderation and Trust and Safety teams.

¹¹⁷ Charles Coxhead, “A Guide to DMCA Protections, Procedures, and Takedowns,” WP DevShed (2017), <https://wpdevshed.com/dmca-guide/>.

¹¹⁸ Interview with Craig Colgan (pseudonym), former employee at Facebook.

¹¹⁹ See also Yoav Dror, “We are not here for the money’: Founders’ manifestos,” *New Media & Society*, 17(4) (2015): 540-555.

¹²⁰ Interview with Craig Colgan (pseudonym), former employee at Facebook.

¹²¹ Interview with Evan Engstrom, executive director at Engine.

¹²² Interview with Michael Hargrove (pseudonym), an official from a U.S. Federal agency. Also see (citation from 230 papers).

At the same time, the law places a low threshold on user base to be subject to these regulations, setting it at two million users, and also contains ambiguous language that can be difficult for platforms to interpret.¹²³ Despite requests that US law provide more clear guidelines, the platforms we spoke with warned that the high fines for violations in Germany will lead to more pre-emptive takedowns, and the use of automated flagging and removal, despite these technologies not necessarily being attuned to sensitivities in cultural context or varying linguistic cues.¹²⁴ Other companies we spoke with noted that they were not a “social media company” or that they had not yet reached the registration numbers to be worried about the law. For a law that’s been colloquially referred to as the “Facebook Law,”¹²⁵ being *not-Facebook* and not industrial becomes suddenly relevant.

Additional disinformation or “fake news” laws enacted by countries like Malaysia are sparking concerns that such rules will be used to enact censorship. According to reports from Reuters, the Malaysia Anti-Fake News Act has already been used to convict a Danish citizen over inaccurate criticism of police over social media.¹²⁶ Reports from *The Guardian* note how authoritarian leaders have co-opted President Trump’s use of the term “fake news” to criticize media, and there is concern that laws enacted by regimes with histories of human rights abuses will use these laws to quell dissidents. Concerns about the way criticism of media is happening within the United States by government actors has led to similar concerns about how credibility and authority of information could be politicized if integrated formally into the law. And yet, a lack of regulation around hate speech and disinformation is frequently being used by authoritarian governments as well, who may be benefitting from confusion and violence stemming from false and inflammatory content spreading online (as has been suggested is the case in Myanmar).¹²⁷ Governments concerned that social media posting may be sparking violence within their country have blocked specific sites entirely, or at least for short periods of time. Authorities in Sri Lanka blocked access to Facebook and WhatsApp for a period of time following violence between Buddhist and Muslim ethnic groups in the country that they argue began following inciting posts on the social media network.¹²⁸ In Papua New Guinea, the Communications Minister, Sam Basil, announced in June 2018 that they would shut Facebook down within the country’s borders for one month to identify and remove fake accounts, study the network’s impact, and potentially even build a “local alternative.”¹²⁹

¹²³ It’s not yet specified as to whether this means two million German users, or users worldwide. This ambiguity was cited by many of our respondents as the reason why they would not be subject to the law, while others were confused as to whether it would apply. As an example of ambiguous language

¹²⁴ Interview with Alex Feerst, Head of Legal at Medium (on takedowns).

¹²⁵ Claudia Haupt, “Online Speech Regulation: A Comparative Perspective,” Presented at the American Political Science Association, August (2018).

¹²⁶ Reuters in Kuala Lumpur, “First person convicted under Malaysia’s fake news law,” *The Guardian* (2018), <https://www.theguardian.com/world/2018/apr/30/first-person-convicted-under-malaysias-fake-news-law>.

¹²⁷ Jonathan Head, “Outlaw or Ignore? How Asia is Fighting ‘Fake News,’” *BBC News* (2018), <http://www.bbc.com/news/world-asia-43637744>.

¹²⁸ Zaheena Rasheed and Amantha Perera, “Did Sri Lanka’s Facebook ban help quell anti-Muslim violence?” *Al Jazeera* (2018), <https://www.aljazeera.com/news/2018/03/sri-lanka-facebook-ban-quell-anti-muslim-violence-180314010521978.html>.

¹²⁹ Al Jazeera News Agencies, “Papua New Guinea to Ban Facebook for a Month,” *Al Jazeera* (2018), <https://www.aljazeera.com/news/2018/05/papua-guinea-ban-facebook-month-180530053406737.html>.

How platforms will treat laws they may disagree with is beyond the scope of this current white paper. However, most of the platforms we spoke with said they respect national sovereignty and will obey the laws of the countries in which they are operating, prioritizing this value above their company's own views as long as the country in question issues takedowns and requests through the appropriate channels. Discovering *what those channels are* is a key part of this process, and having offices, personnel, or experts to draw on *within* every country they are operating, is an important (if not often neglected) component in addressing both national law and the major content concerns emerging from local contexts.¹³⁰ Part of this would entail addressing differences in how companies organize their policy teams and encouraging more transparency and oversight into how personnel interact with local governments, and, more importantly, civil society, and watchdog organizations could be one step in the direction toward situating policy decisions regarding hate speech and disinformation within ground-level cultural and political dynamics.

¹³⁰ This was cited in many of our interviews with platform stakeholders.

CONCLUSION:

Sizing Up Platform Regulation

Where organizations of different sizes seek to balance whether they can be contextual with being consistent, there are several additional differences in how they create and enforce policy. As noted, online speech platforms typically differ in the degree to which they separate policy development and enforcement. Smaller-scale organizations typically house policy development and enforcement closely together, whereas the industrial approach typically separates them both organizationally and geographically. Community-reliant organizations draw clearer lines on what type of content (or rather, conduct) is acceptable and where they are likely to intervene, thus leaving the bulk of policy development and enforcement to their volunteer moderators.

This paper contends that understanding *how* platforms make moderation decisions, and where they share challenges (or diverge), is one step toward determining **how to design more nuanced solutions.**

As we become more aware of the role private platforms play in regulating speech, it becomes necessary to create more avenues to not only oversee this decision-making, but create mechanisms to redress the impact these rules (or lack thereof) are having, particularly on marginalized communities around the world. At the same time, public oversight needs to incorporate not only the context of speech, but the organizational dynamics of platforms, to understand where new rules should be developed (for types of content), and where more resources are necessary. This white paper has examined the challenges facing content moderation in teams of all sizes as they attempt to draw clear lines around acceptable or unacceptable content at a scale that frequently transcends geographic borders and blurs cultural norms. This paper contends that understanding *how* platforms make moderation decisions, and where they share challenges (or diverge), is one step toward determining how to design more nuanced solutions. The different organizational dynamics of these content moderation teams represent different values guiding their policies and practices, as well as the relative stage of the team itself. Artisanal models enable a contextual approach to moderation but are constrained in terms of applying these rules consistently, which makes them vulnerable as

these companies grow. As companies formalize into industrial models, rules that must take into account different cultural context needs become too rigid; not able to take into account cultural differences such as what is *newsworthy*, *hate speech*, or *disinformation* within a specific region. In this sense, even larger, well-resourced platforms can have serious linguistic and cultural gaps which limit their capacity to respond in ways that satisfy growing public concerns.

As policymakers within these companies try to draw lines around the kind of content they want or do not want on their platforms, they become less the “arbiters of truth” than the arbiters of hate, arbiters of harassment, and arbiters of disinformation around the world.

Platforms, even the smaller ones, have been given an immense amount of power and responsibility within the information era. Like other communication systems that existed before – newspapers, broadcast, and cable – they have the capacity to shape what we see and do not see and the capacity to personalize content and advertise to individuals, at a granularity never before seen. Though they provide more avenues for individuals to communicate, that has come with its own kind of liability, both financially and reputationally. As policymakers within these companies try to draw lines around the kind of content they want or do not want on their platforms, they become less the “arbiters of truth” than the *arbiters of hate, arbiters of harassment, and arbiters of disinformation* around the world. These decisions are not easy, and drawing lines around some speech one person finds objectionable *almost always* has unintended consequences for other speech. The policymakers at these platforms, who are in charge of either creating or defending content rules, are not elected to these positions. Though as Klonick notes, these individuals have a powerful role in governing speech; the individuals making policy are neither judges nor juries. They aren’t even always lawyers (although some are); some may be activists or advocates for a certain political position. More likely, most are just showing up to jobs with an immense amount of public responsibility and little external guidance (and not enough internal resources) on how to make these decisions beyond the bubbling up of public opinion coming either from news media or from coordinated efforts on their own networks. Though public outrage against figures such as Alex Jones, who was allowed to remain on a wide array of platforms despite obvious violations to community guidelines, is occasionally an effective strategy for removing objectionable content from sites, it is neither feasible at scale, nor fair for countries as *impacted by hateful rhetoric as is the United States*,¹³¹ but with a public that has significantly less influence over these platforms.

¹³¹ Jack Nicas, “Alex Jones Said Bans Would Strengthen Him. He Was Wrong,” The New York Times (2018), <https://www.nytimes.com/2018/09/04/technology/alex-jones-infowars-bans-traffic.html>.

As platform companies increasingly try to find the “right” answer to policing difficult content types, it is difficult to imagine how they will achieve the results that will satisfy the diverse communities that are now reliant upon on them. Within the United States, this issue has become increasingly complicated. Calls for the regulation of platforms as “public goods”¹³² or utilities are not new, and they’ve typically been adopted by politicians on the left, like Elizabeth Warren or Bernie Sanders, in reference to the growing power and market share of platforms.¹³³ Recently, however, this discourse has been picked up by unlikely candidates – far-right-wing politicians and media commentators – who have recently been brandishing the threat of regulation loudly.¹³⁴ These calls have often been made alongside critiques of platforms’ moderation programs or algorithmic prioritization, perceiving their far-right rhetoric is being overly censored by platforms with policies against disinformation and hate speech (that has in no way been proven).¹³⁵ Though it is unlikely the current administration will be successful in its quest to regulate platforms such as Google – as noted above, courts have largely interpreted Section 230 of the Communications Decency Act generously in favor of immunity for platforms¹³⁶ – it still leaves us with the question of *how exactly we make the bureaucracies of content moderation more participatory and democratic, regardless of politics.*

This is also not just an issue for the United States, making the question *who should make and enforce* content rules even more complicated. Though this white paper places a large focus not only on the cultural environments in the United States and in Western Europe, the content platforms we have been discussing are having undeniable consequences for speech and the safety of individuals all around the world. At this point, it has become a question of whose laws or norms will prevail and become adapted into algorithms and content moderation programs or, conversely, how platforms will work to create the capacity within each country (and each region within each country) to achieve the level of specificity and context needed to not only obey the law (when it exists) but to also track and address hate speech and disinformation within that language and culture. At the

It still leaves us with the question of how exactly we make the bureaucracies of content moderation more participatory and democratic, regardless of politics.

¹³² Mark Andrejevic, “Public Service Media Utilities: Rethinking Search Engine and Social Networking as Public Goods,” *Media International Australia* 146 (1)(2013): 123-132.

¹³³ Elizabeth Warren, “Reigniting Competition in the American Economy,” Keynote Remarks at New America’s Open Markets Program Event (2016), https://www.warren.senate.gov/files/documents/2016-6-29_Warren_Antitrust_Speech.pdf.

¹³⁴ For an example, see positive coverage of Tucker Carlson on Breitbart, “Google Should be Regulated Like the Public Utility It Is.” <https://www.breitbart.com/video/2017/08/15/tucker-carlson-google-regulated-like-public-utility/>

¹³⁵ Nancy Scola and Ashley Gold, “How Trump could hurt Google,” *Politico.eu* (2018), <https://www.politico.eu/article/how-donald-trump-could-hurt-google/>.

¹³⁶ Danielle Keats Citron and Benjamin Wittes, “The Problem Isn’t Just Backpage: Revising Section 230 Immunity,” *Georgetown Law Technology Review*, 453(2018).

same time, platforms must scale up these efforts so quickly in response to emerging regulations that they may feel tempted to adopt automated detection technologies that dissolve differences between communities and search for easy, formulaic rules to address complex concerns. The major concern in an era of convergence is and will always be *context* — providing more of it to users as they consume information, more to moderators as they assess content, and more to stakeholders both internal and external to platform companies as they make decisions about content policy. Understanding the differences in approach may help impose rules that can help each type of platform do this best.

ACKNOWLEDGMENTS

The author would like to express her appreciation to the individuals who generously reviewed this report and gave critical feedback and insights: Tarleton Gillespie, Philip M. Napoli, Claudia Haupt, and Chris Tenove. danah boyd was pivotal in guiding this report at every moment; the initial idea stage, during early interviews, and then honing arguments in later stages. Janet Haven, Matt Goerzen, and Andrew Selbst were gentle sounding boards for many of the initial findings. Patrick Davison provided astute and thoughtful editorial guidance, and an open mind (and calendar) to talk through many of these ideas. The entire team at Data & Society supported this project in innumerable ways. We would also like to thank our 30+ interviewees across media, civil society, platform companies, and government, for their time, energy, and insights on critical issues and concerns.

This report benefitted from feedback at the Stanford Workshop for Digital Technologies and Democratic Theory, hosted by the Center on Philanthropy and Civil Society.

For further correspondence about this report please contact Robyn Caplan at robyn@datasociety.net.

Data & Society

Data & Society is an independent nonprofit research institute that advances new frames for understanding the implications of data-centric and automated technology. We conduct research and build the field of actors to ensure that knowledge guides debate, decision-making, and technical choices.

datasociety.net

@datasociety

Design by: Andrea Carrillo Iglesias, Rona Binay