Yale Journal of Law & Technology, Volume 23

&

The Justice Collaboratory at Yale Law School

Present:

IN A NEW LIGHT:
SOCIAL MEDIA GOVERNANCE RECONSIDERED

Sudhir Venkatesh, Tom Tyler, Tracey Meares & Farzaneh Badiei[*]

The ubiquity with which platforms for online interaction have arisen and spread across the world has kept private companies, governments and the people using these platforms playing continual catch-up, trying to both utilize the new possibilities created by internet-based communication and protect users from both traditional and newly emerging harms that occur when interacting with others. Many of the problems emerging in online platforms mirror long-term issues associated with governing interactions in real world communities, while some are unique to the new internet world. Three types of governance are important. One is self-governance, the ability of users to cooperate with others to manage their own online interactions. A second is platform governance, the capacity of private vendors to effectively manage what occurs on their platforms. Finally, online communities may cross political boundaries but they exist within a complex matrix of local, national and international regulatory communities. These all play some role in governing the form and content of online platforms.

[*] Sudhir Venkatesh, Williams B. Ransford Professor of Sociology, Columbia University; Tom Tyler, Macklin Fleming Professor of Law and Professor of Psychology and Founding Director of The Justice Collaboratory, Yale Law School; Tracey Meares, Walton Hale Hamilton Professor of Law and Founding Director of The Justice Collaboratory, Yale Law School; Farzaneh Badiei, Director of Social Media Governance Initiative, The Justice Collaboratory, Yale Law School.

Our particular concern is with platform governance of these spaces for online interaction. Most platforms originated by conceptualizing themselves as pass-through architecture for interpersonal communications. Their creators no more imagined the prospect of regularly reading people's messages than the post office workers would imagine reading people's paper letters. Moreover, platform creators viewed their role as facilitating positive social communications among willing participants. And the rise of platforms for online interaction has facilitated traditional social communications, enabled people to make new connections and helped to maintain connections in better ways. Our social world has moved from the letter to the telephone to the Tweet or post. These new forms offer an unparalleled capacity for rapid and personalized connections across broad distances. Platforms *have* facilitated positive social communications among willing participants.

Of course, as more of our social world occurs online, the problems that plague the off-line social world follow. People can use online communications to threaten, bully and embarrass others in particularly effective ways. They can use internet platforms to push out negative messages about social and political issues, messages ranging from racism to hate speech and even advocating support for terrorism. The same tools that help people make new friends and form communities around a shared interest in gardening also enable extremists to recruit new members. The proliferation of negative content has forced platforms to become content regulators, whether or not they want to take on that role. In some cases, existing problems in the off-line world are not just perpetuated but, rather, intensified online. Algorithms, core to the technical infrastructure and scalability of these platforms, are prime examples of this phenomena. Widely used, many algorithms are aimed at mimicking human decision-making for efficiency and scalability's sake. Most often, these algorithms reinforce systematic biases of the individuals and organizations training, building, and deploying them. At worst, feedback loops in algorithms can inadvertently magnify these biases further marginalizing individuals or groups.

Many platforms have looked to the deterrence model common in legal settings as an initial framework through which to regulate content. Policy teams create rules and platforms create technical and operational mechanisms to evaluate user content against those rules. Those who violate rules by posting violating content get sanctioned in some way, typically with a graduated series of sanctions. Users' posts are removed, their accounts might be suspended for some period of time, or users might even be banned from a platform. In adopting this approach online platforms have inherited both the strengths and weaknesses of traditional law.

Studies show that in democratic societies like the United States, deterrence models work to change behavior, although not particularly well. Low-level offending presents an especially challenging environment for such models, a situation typical of online platforms. On the other hand, online platforms have notable advantages over real world legal authorities because they can more readily scan user platform behavior for rule conformity and have much greater control over when and how users can utilize the platforms. Still, problems like those faced by legal authorities arise. Some are related to defining and implementing rules for content moderation, which involves turning abstract ideas into practical and operational review guidelines used by a global workforce of agents reviewing vast amounts of content for violations of these rules. Since users imagine that their communications move more or less immediately to their intended audience, platforms have sought rapid algorithms to detect harmful content, moving the initial problem of flagging problematic content evaluation from human to machine.

Human review often follows flagging by machine algorithms, but that process takes time. Human review also occurs in response to people's complaints, so harmful content may be viewed by many users prior to any platform action. Platform owners, since they control access to their platform, can also more successfully sanction offenders than can real-world legal authorities. Here too, however, users can seek to evade sanctions or bans by using multiple accounts or moving to private sites.

Platform content, especially content that violates content moderation rules, is continually in the news, reflecting limitations in the existing governance models for content moderation. On the other hand, the newsworthiness of apparent content moderation failure may simply reflect the centrality that social media has assumed in people's social interactions.

These newsworthy moderation challenges also reflect a lack of consensus about what problematic content is and how to address it. On the one hand, there are calls for flagging or taking down material that some groups feel is problematic. At the same time, others complain about the suppression or exclusion of that same content they regard as valuable. What is desirable and what should be flagged or even banned depends upon underlying values and is an active debate. While this issue conflicts particularly with political speech, even efforts to limit nudity encounter differences in people's values about what forms of nudity are and are not offensive.

Regardless of their reasons, many people are dedicated to thinking through better governance models of online platforms. Here, a multidisciplinary group of researchers reconsider the issues involved in this rapidly evolving space and consider new ideas and alternative possibilities for social media governance. This issue brings together a group of prominent scholars using a broad array methods and theoretical perspectives to address platform governance in a new light and in an evidence-informed fashion.

Our aim for this special issue is to bring a few novel approaches to platform governance which can be applicable to social media and other online platforms. The different scholars included in this issue approach social media governance through different lenses, and sometimes use different terminology (e.g., "platforms" vs. "technology firms" vs. "social media companies"). Yet the common thread is the importance of exploring new ideas for managing the social impact, good and bad, that these large players have in our society. Our hope is that this issue will spur as lively a conversation about these topics as we had at the mini conference at which each of these papers was presented. These papers reflect not only the ideas of their authors but also the feedback from the distinguished group of scholars convened to comment upon them. To make progress upon these ideas we will need a dedicated cohort of people willing to think about these problems in a different way. This issue represents our effort to create such a group.

Rethinking Models of Social Media Governance

As noted, many platforms have reacted to the problems of negative content by trying to engage in some form of content moderation. This involves identifying problematic content ranging from nudity to hate speech. A review of both rules and strategies to enforce them reveals that platforms use the legal model of suppressing bad behavior through the threat or use of sanctions. Badiei, Meares & Tyler argue that this is a mistake. Platforms should encourage users to voluntarily internalize the rules and willingly follow rules and engage in positive behavior and healthy interactions. The key to this model is to change what users want to do and thereby discourage the emergence of bad behavior in the first place. This argument has two parts. The first mirrors recent reform efforts in criminal justice in recognizing that when people view rules and authorities as legitimate, they feel a responsibility to follow those rules and authorities. This strategy promotes rule adherence in a way that lessens the need for surveillance and sanctioning. It is especially important in an arena like online platforms in which most users are well intentioned and many rule violations come through a lack of awareness of the rules.

A legitimacy-based model has the second advantage of building identification with other people in the community, leading users to want to make their online communications positive, facilitating healthy interactions and vital online communities. Evidence demonstrates that it is possible to create online platforms that promote user identification with their communities and which enhance the legitimacy of platforms and of their regulatory efforts.

In a similar vein, Schoenebeck and Blackwell argue that social media platforms have often followed the traditional legal system in focusing on punishing offenders, without paying attention to how to mitigate conflicts or repair harm to victims. Social media platforms are punitive rather than reparative and focus on removing harmful content or users. They neglect the task of helping the victims of the abuse. These authors argue that platforms would benefit from adopting reparative approaches centered on global values such as dignity, accountability, and community. Although negative content may not be illegal, it still harms others, and platforms should adopt a broader perspective which recognizes the desirability of focusing on the well-being of those who have experienced negative online interactions.

Policies and Practices for Content Moderation

Although several contributors argue that platforms for online interaction overemphasize content moderation, content moderation still is necessary, so one must ask how can moderation best be achieved? Companies struggle to find ways to implement their desired goal of lessening or even eliminating exposure to "bad" content. They are trying to find ways to identify content that would be generally viewed as bad content. One of the more challenging examples of this struggle is found in the arena of politically or socially controversial content. Here there is often disagreement about what type of messaging is inappropriate and who should make such decisions. One approach that some platforms have used is not to remove content but to give it less priority in user feeds. Another approach, discussed by Wihbey, et al, is to post content but provide some type of warning or explanation, a practice called labelling. Such labelling can take different forms. It might involve an effort to correct factual errors and aim against misinformation. It can also be motivated by a desire to help users recognize alternative perspectives on a particular issue, including perspectives that are the opposite of their own or that are held by "experts" on a topic. Labels can encourage readers to read supplementary material that the platform believes clarifies or even contradicts a particular message.

Wihbey et al. analyze this specific governance method of "labeling." They do so with an epistemological approach. Their argument is that, despite the promise of labeling as a strategy, it has thus far been mostly tactical, reactive, and without strategic underpinnings. Wihbey et al. argue that social media companies have been struggling to devise and implement policies on handling misinformation that the public finds generally palatable. In place of consistently-enforced policies that are transparent to all parties, large platforms such as Twitter and Facebook have been responding to different instances of misinformation in a seemingly piecemeal fashion: downranking some posts, removing others, and labeling or "fact-checking" still others. This approach has led to social blowback, especially in those cases where algorithms are involved. They therefore argue against defining success as merely curbing misinformation spread. The healthy way of labeling is to consider it from an epistemic perspective and to take the "social" dimension of online social networks as a starting point. The strategy in this article emphasizes how the moderation system needs to improve the epistemic position and relationships of platform users—i.e., their ability to make good judgments about the sources and quality of the information with which users interact on the platform—while also respecting sources, seekers, and subjects of information.

Obviously, in order to govern online platforms by moderating content it is necessary to have criteria that define good and bad content. Often people feel that bad content is self-evident. For example, Supreme Court Justice Potter Stewart defined the Court's standards for obscenity by saying "I know it when I see it." Online platforms, in contrast, have developed elaborate codebooks for their human reviewers and have tried to develop computer programs which embody the same rules. This requires a two-step process. First, identifying principles (e.g., "no nudity"). Those rules then have to be elaborated into guidelines that are specific enough that they can be utilized by either a human coder or an algorithm.

Pineda is concerned with the origin of the principles and, in particular, with the question of whether there are any universal principles that can rise above the values of any particular society or culture. Social media platforms began in America and they sometimes employ general principles derived from America to determine their rules. Even if this were reasonable, the rise of alternative platforms in other societies makes this approach unrealistic. So where will standards come from in the future?

Pineda argues that we can best analyze the challenges of content governance by understanding the debates and conversations that take place about culture, cultural relativism, and the universality of human rights. In particular is the West imposing its values on everyone in the guise of "universal values"? How can we resolve this through anthropological means? The ongoing work of formulating "universal" content moderation policies will benefit from understanding the histories and debates in anthropology about cultural relativism and human rights universalism in order to avoid some of the pitfalls that are inherent in this kind of global governance. Anthropology can help us distinguish between values that are universal amid the difference in the expression of values across the world. Just like the universality of human rights has been scrutinized in global governance, the general standards that social media platforms have asserted have been contested.

Credibility Online: Who Do We Trust?

Social scientists have long argued that the willingness to trust other people is central to engaging in exchanges with others. Such exchanges frequently require people to take risks based upon the belief that the other people involved in an interaction have benevolent and sincere motivations and are not seeking to take advantage of them. People have developed strategies for evaluating the trustworthiness of others in real-world interactions. However, there are questions about the degree to which a similar level of trust can be established and maintained remotely, an issue central to rapidly emerging online platforms. The core question is whether a participant in an online market like eBay is willing to trust another in the same way that people have trusted others in their community in the past, and, as in real world interactions, what mechanisms can be identified to facilitate such trust and make online markets viable.

Parigi and Lainer-Vos argue that the rise of two-sided online markets and the centrality of reputation systems have undermined trust. Instead of trust being a byproduct of interpersonal interaction, thin trust in online markets demands methodical cultivation of trust in a mostly impersonal and domain-specific fashion.

Trust is central to exchange and cooperation. In offline situations people continually struggle to decide whom to trust and when to take risks by being vulnerable to others. If people never take risks, they gain little from being in markets. If people trust too uncritically, they may rely on others who do not keep their promises. Traditional discussions of trust emphasize the role of reputations in enabling trust. Someone who might break another's trust in one situation recognizes that if they acquire a reputation for being untrustworthy no one will exchange with them in the future. Reputations in traditional communities were a shared property, and people sought out and interacted with trustworthy others. Parigi and Lainer-Vos argue that the online world poses challenges for people trying to determine whether to trust someone else. Consequently, the nature of trust is changing in this new domain.

Future Research in Online Governance

This special issue represents our attempt to contribute to this growing need for rethinking online platform governance. Undoubtedly, we will continue this work through a network of interdisciplinary scholars within the Justice Collaboratory's Social Media Governance Initiative. As we think through what future research could contribute to this conversation, it is important to highlight some areas we are particularly concerned about, like shifting the focus of scholars and policy-makers towards the design, architecture, and infrastructure decisions that shape governance.

If the prevailing model of content moderation is not the most desirable way to manage platforms, a key question is why this model exists and how it was built. At the center of the organizational culture of most online platforms is the product group. This is the group that manages the architecture of the platform: many hundreds of engineers, designers, and product managers. Because this group dominates these companies, the issue of content moderation within these organizations has been generally viewed as a technical one, something amenable to management through simple screening algorithms that can detect and remove nudity or hate speech.

The insight that content moderation is viewed as a technical problem within the purview of product teams helps to illuminate why external regulation efforts have been problematic. External constituencies typically interface with the legal and managerial elements of online platform companies—typically policy teams rather than these product teams. This means that both scholars and those seeking platform changes rarely look at product design culture and how it shapes content moderation in technology firms. The fact that content governance is housed in product units reflects the history of the evolution of platforms, which was focused on solving technical problems, not addressing issues complex social issues of content acceptability across the globe. Some of the recent efforts to share data with scholars through Transparency Reports or create Oversight Boards are examples in which the corporate leadership draws energy away from product divisions that have more substantial impact on governance of platform users.

As more public attention is paid to the impact of social media and other internet companies, it would be worthwhile for outsiders to redirect some their efforts toward the technology creation efforts of product teams. As we look towards furthering the conversation over platform governance, we need to spend more time thinking about platform architecture and the design of infrastructure in addition to the current focus on the rules themselves. Safety in automobiles can be a very helpful analogy in this regard. While speed limits and other rules of the roads are important to ensure public safety, far more critical in saving lives are the design of the cars we drive—airbags, crumple zones, or seatbelts—and infrastructure of the roads we drive on—rumble strips, clear signage, or banked turns.

This discussion also highlights the issue of platform motivations. Newspapers struggle with the problem that sensational news sells papers. In the same way, online platforms are for-profit entities. Their profits flow from putting ads in front of their users, selling knowledge harvested about its users to advertisers allowing vendors to target likely candidates for their products. This means that if extreme or salacious content attracts attention, it is to the benefit of the company to highlight such content in order to attract and retain the attention of their users. Content moderation is in conflict with this business model. As a consequence, it is sometimes difficult to discern whether companies are actually interested in effectively moderating such content or are interested in presenting an image of civil responsibility that can fend of government regulation, oversight and organized consumer push back. Discerning the internal dynamics of organizations running online platforms is also important in future study of online governance.

CONCLUSION

We are hopeful that this material contributes to the debate about how humanity might govern itself online. These papers demonstrate how to apply interdisciplinary approaches to social platform governance and go beyond the currently dominant governance mechanisms which this group collectively argues have not so far been effective. We believe the papers provide an important contribution to the technology governance landscape and we thank the editorial board at the *Yale Journal of Law and Technology* for their collaboration in publishing this special issue.

Table of Contents

COMMUNITY VITALITY AS A THEORY OF GOVERNANCE FOR ONLINE INTERACTION

Farzaneh Badiei, Tracey Meares & Tom Tyler

OVERVIEW

Governance of platforms for online interaction has targeted primarily what users themselves put up online. That is, platform governance mechanisms typically focus on managing problematic content ranging from nudity to hate speech, something that platforms call "content moderation."[1] A review of both rules and strategies to enforce them reveals that moderation is focused on identifying and punishing bad behavior.[2] We think this is a mistake for at least two reasons. First, framing the issue of online content moderation primarily as an effort to find and suppress undesirable actions as opposed to focusing on strategies to encourage users to voluntarily internalize rules and engage in "good" behavior replicates the mistakes the criminal justice system has made in managing behavior identified as criminal "in the real world." Second, and perhaps more important, focusing on identifying and punishing bad behavior prioritizes elimination of bad behavior over the creation of a framework that facilitates healthful interaction. Such healthy interaction discourages the emergence of the bad behavior in the first place. We argue that just as in the real world it is better to facilitate and encourage healthful community interaction to avoid crime, platforms should engage in the project of creating infrastructures to encourage strong, healthful communities online.

[1] Robyn Caplan, *Data & Society, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches* (2018), https://datasociety.net/library/content-or-context-moderation/.
[2] Tom Tyler et al., *Social Media Governance: Can Social Media Companies Motivate Voluntary Rule Following Behavior Among Their Users?*, 17 J. EXPERIMENTAL CRIMINOLOGY 109 (2019).

These communities are more likely to be self-regulating communities that need less external policing. In this article, we discuss how to import these ideas into the governance structure of platforms for online interactions.

We rely on our work with respect to the operation of the criminal justice system in the real world to demonstrate that organizing governance structures around the social psychology of procedural justice can produce positive results regarding voluntary compliance with rules and laws. People respond positively to procedural justice in the criminal justice system, in contrast to deterrence approaches premised upon the notion that people comply with rules and laws because they fear the consequences of failing to do so. Procedural justice strategies treat individuals as engaged agents who should have a part to play in the overall fair functioning of the system. Our experience working with platforms demonstrates that many rely on deterrence-based "get-tough" strategies to achieve compliance, and there is little reason to believe that such approaches work any better for social media than for criminal justice. Using procedural justice strategies to shape people's behavior online might prove useful.

Our larger point is that creating vital communities should be a key goal of governance as a general matter, whether on- or offline. We can divide online vital communities into two groups: those that engage in constructive and positive interactions on online platforms and those that use the opportunities provided by the platforms to manage their offline issues. An example of the former is the ongoing discussions that happen in Reddit or Facebook groups about race in the United States, stimulated by Black Lives Matter. An example of the latter is the efforts of online community groups to manage COVID-related problems in their communities—for example, Nextdoor's groups that offer help to the elderly and others in need during the pandemic.[3] In both cases, the goal should be to leverage the possibilities of online platform communication to enhance community vitality and well-being. This involves lowering the level of negative or divisive online interaction and raising the constructive and problem-solving communication.

Building on theory and research demonstrating that people care more about the procedural fairness with which decision-makers treat them than the outcomes themselves, we explain how procedurally just treatment can encourage online community members to voluntarily follow platform rules and work constructively with each other to solve problems.

The first goal of this approach is to build self-regulatory models of content moderation. To do so, it is important to create commitment to rule-following that involves users' sense of obligation rather than their concerns about punitive measures, like having their posts or accounts blocked or permanently banned. To the degree that this model is effective, it is not necessary for platforms to try to identify wrongdoing. People more willingly follow the rules when they self-moderate than when coerced to take certain actions.

This approach has a second goal of building community vitality. The core of our argument is that the absence of harm is not the same thing as the presence of vitality. Community vitality is present when there are high levels of economic prosperity, social capital and well-being.

The goal of the suppression of harmful content may be a necessary beginning, but it is important to ask whether the strategies

[3] *Using Nextdoor to support your neighborhood during the COVID-19 pandemic*, NEXTDOOR, https://help.nextdoor.com/s/article/Using-Nextdoor-to-support-your-neighborhood-during-this-crisis? (last visited Feb. 15, 2021).

being used contribute to a long-term goal of building a vital community. If social media platforms adhere to procedural justice in their design, in their moderation efforts, and in their decision-making processes around the structuring of online groups, we believe that they can enhance community vitality and cooperation among online users. Those users work more constructively together, build shared identification and solidarity, and reach consensus approaches about how to address the issues that concern them.

INTRODUCTION

It was only in 2018 that Facebook first made public the guidelines that its moderators used to enforce its community standards.[4] This is not to say there had never been any rules. For twelve years, Facebook had prohibited publishing many types of objectionable content on its platform.[5] Despite the existence of these rules, however, users technically had no idea what the rules were or how Facebook enforced them unless they happened to violate them. When users did violate a Facebook content moderation rule, they were punished by being banned from using the platform for a particular period of time. For many years, Facebook unilaterally took content down, and there was no opportunity for the users to engage with the company about the consequence of a violation. Facebook's historical approach created a dynamic that remains a pillar of the relationship between social media platforms and their users: the platform imposes and enforces rules and the users obey.[6]

[4] Monika Bickert, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, FACEBOOK (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/.

[5] Facebook, as early as 2006, had rules that governed the users and their content. *Member Content Posted on the Site*, FACEBOOK (Jan. 11, 2006), https://web.archive.org/web/20060118020625/https://www.facebook.com/terms.php.

[6] We do not claim that Facebook did not try to make its policy more community-oriented. We argue that it did not select a way that would involve the community

To achieve prosocial governance and a vital online community the relationship between users and platforms must shift from one focused on platform authority and user obedience to an environment focused on user internalization of rules regarding content and its moderation. In the case of voluntary rule-following, the traditional model encourages people to hide their behavior and requires authorities to search for rule-breaking. This is a challenging task. When people are invested in following the rules, they view doing so as a personal obligation and do it without reference to whether their conduct can be observed and sanctioned.[7] While online platforms may seem adept at monitoring their users' behavior, in reality, they have found that people find creative ways to hide.[8] For example, they may create multiple accounts and fake identities. Social platforms have inevitably been thrown into a role familiar to police departments: watching their community and trying to identify violations.

of users effectively and create a more bottom-up approach. Also despite its efforts over the years to publish its policies, the implementation of those policies remained obscure for the day-to-day user. For the changes in Facebook content governance approach over the years, see Rotem Medzi, *Enhanced self-regulation: The case of Facebook's content governance*, NEW MEDIA & SOC'Y (2021). In 2009, Facebook announced that it planned to try new forms of governance and giving more authority to the users. The problem with that approach, which later failed, was that it did not allow the users to self-regulate, and it was not very clear how Facebook enforced those policies. It only allowed the users to vote on the changes that Facebook had undertaken. *Facebook Opens Governance of Service and Policy Process to Users,* FACEBOOK (Feb. 26, 2009), https://about.fb.com/news/2009/02/facebook-opens-governance-of-service-and-policy-process-to-users/. This new governance approach failed since not many participated in voting. *Results of the Inaugural Facebook Site Governance Vote*, FACEBOOK (Apr. 24, 2009), https://web.archive.org/web/20090430215524/http://blog.facebook.com/blog.php?post=79146552130. In 2012, Facebook decided to remove the voting mechanism altogether and instead reach out to a select number of third-party experts. Elliot Schrage, *Proposed Updates to our Governing Documents*, FACEBOOK (Nov. 21, 2012), https://about.fb.com/news/2012/11/proposed-updates-to-our-governing-documents/.

[7] *See* TOM R. TYLER, WHY PEOPLE OBEY THE LAW (2006).

[8] Lauren Reichart Smith et al., *Follow Me, What's the Harm: Considerations of Catfishing and Utilizing Fake Online Personas on Social Media*, 27 J. LEGAL ASPECTS SPORT 32 (2017).

This approach to motivating rule compliance should be familiar to anyone who works in the criminal justice system or understands how it works. It is an approach based upon the idea that people will follow rules or laws because they fear the consequences of failing to do so and that in order to ensure that people do follow rules, the punishment or the threat of punishment must be severe enough to motivate a rational actor to follow the rules. Two of us, Professor Meares and Professor Tyler, have spent the past two decades explaining the ways in which this approach to compliance in criminal law does not work well and often in fact undermines the stated goals of the system.[9] We have argued in favor of approaches that encourage internalization of rules based on enhancing citizen trust in legitimacy of various kinds of authorities.[10] We characterize these approaches as prosocial in that their goal is to promote and enhance existing positive norms of behavior as opposed to making central the ferreting out and punishing of bad behavior. In this paper, we apply ideas we have developed in the criminal justice space to online platforms,[11] and we theorize that prosocial governance

[9] TYLER, *supra* note 7; TOM R. TYLER, WHY PEOPLE COOPERATE: THE ROLE OF SOCIAL MOTIVATIONS (2013) [hereinafter WHY COOPERATE]; Tom R. Tyler, *Enhancing Police Legitimacy*, 593 ANNALS AM. ACAD. POL. & SOC. SCI. 84 (2004) [hereinafter *Police Legitimacy*]; Tom R. Tyler & Tracey L. Meares, *Procedural Justice Policing*, *in* POLICE INNOVATION: CONTRASTING PERSPECTIVES 71 (David Weisburd & Anthony Braga eds., 2019); Tracey L. Meares, *The Path Forward: Improving the Dynamics of Community–Police Relationships to Achieve Effective Law Enforcement Policies*, 117 COLUM. L. REV. 1355 (2017); MEGAN QUATTLEBAUM ET AL., JUSTICE COLLABORATORY AT YALE L. SCH., PRINCIPLES OF PROCEDURALLY JUST POLICING (2018); Tracey L. Meares et al., *Lawful or Fair? How Cops and Laypeople Perceive Good Policing*, 105 J. CRIM. L. & CRIMINOLOGY 297 (2015); TOM R. TYLER, LEGITIMACY AND CRIMINAL JUSTICE: AN INTERNATIONAL PERSPECTIVE (2007); Tom R. Tyler, *What Is Procedural Justice—Criteria Used by Citizens to Assess the Fairness of Legal Procedures*, 22 LAW & SOC'Y REV. (1988) [hereinafter *Procedural Justice*].

[10] Tracey L. Meares & Tom R. Tyler, *Justice Sotomayor and the Jurisprudence of Procedural Justice*, 123 YALE L.J. F. 525 (2014).

[11] In what follows, we frequently refer to "platforms." Platforms are a means by which people can engage with one another through a network, including the

approaches can contribute to community vitality online. We think that prosocial governance approaches encourage people to follow platform rules and internalize rule-following, and more importantly, to engage with problems and cooperate to solve them constructively. Both of these goals are important, though they may have different ends. The first is good for platforms and their operation. The second is good for society. Since these are complementary ends, we treat them as equal goals.

Platforms must also engender cooperative engagement among community members who use the platform as a forum to address common problems constructively. Community members' motivations for creating a cooperative space are several. First, platforms want their users to enjoy their time on the platform and find it both a positive experience and one that is useful to them in managing the problems in their lives. This cooperative engagement is also important because it enhances the capacity of communities to work together and thereby improves social, economic, and political well-being. When people communicate in positive and constructive ways, they are better able to work together to address common issues and problems.[12] When people are better able to work together, they create stronger communities because they can and do address the needs in those communities more effectively.

Internet. Platforms are always controlled by a single entity. In particular, the function of the platform is subject entirely to the control of the platform operator, so to get access to the functionality offered by the platform, the users must accept the platform's terms of service. Platforms are usually accessed through the World Wide Web but need not be. This use of "platforms" does not include software development platforms or other such uses common in the tech industry; it is primarily societally defined and comprises at least everything popularly described to be a "social media platform," but also includes systems that are often not thought of as social media (such as GitHub or Stack Overflow).

[12] Tom R. Tyler & Steven L. Blader, *The Group Engagement Model: Procedural Justice, Social Identity, and Cooperative Behavior*, 7 PERSONALITY & SOC. PSYCHOL. REV. 353 (2003).

Just as the criminal justice system is a way of governing human interaction in the offline world, there are ways of governing human interaction in the online world.[13] Some of those governance methods depend on similar deterrence strategies to those that have been used in the criminal justice system,[14] so there is reason to believe that reform strategies that apply to the criminal justice system will apply to communities and governance methods online. It is commonplace to use authority-based governance (which depends upon sanctions and deterrence) as opposed to community-based governance (which depends upon willing consent) to manage online behavior.[15] Authority-based governance operates through the rules, practices, and procedures adopted by social media platforms and their employees—decision-makers such as content reviewers, policymakers, and product designers. Authority-based governance is built from the norms and values of the platform and not those of the community it serves.[16] By contrast, in community governance,

[13] A very detailed account of how social media platforms and social networks govern their users can be found in Danah M. Boyd & Nicole B. Ellison, *Social Network Sites: Definition, History, and Scholarship*, 13 J. COMPUT.-MEDIATED COMM. (2007); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2017).

[14] Scholars have mentioned the similarities between platform governance and punitive governance in more detail. We have mentioned these sources in *infra* note 42. To make the similarities more tangible we provide a few punitive approaches here: Tinder (a dating app) permanently bans its users if they violate the rules and is quite unforgiving. Tinder's policy says, "If you violate any of these policies, you might be banned from Tinder. Seriously, don't make us Swipe Left on you—because ***there will be no do-overs once we do.***" *Community Guidelines*, TINDER, https://policies.tinder.com/community-guidelines/intl/en/ (emphasis added) (last visited Mar. 20, 2021). Twitter's enforcement actions are punitive too and can escalate: Twitter can limit tweet visibility and require tweet removal, hide a violating Tweet while awaiting its removal, place an account in read-only mode, and permanently suspend an account. *Our Range of Enforcement Options*, TWITTER, https://help.twitter.com/en/rules-and-policies/enforcement-options (last visited Mar. 20, 2021).

[15] Thomas C. O'Brien et al., *Building Popular Legitimacy with Reconciliatory Gestures and Participation: A Community-Level Model of Authority*, 14 REGUL. & GOVERNANCE 821 (2020).

[16] In Bradford et al.'s transparency report, authority-based governance is presented as top-down governance. In this governance model, the social media

the community—a group of people that share common goals and interests—helps to make and enforce the norms, procedures, and practices by which the platform is governed.

Our goal in this paper is to advance a theory of a self-motivated prosocial production system—that is, a system that by its nature produces a cycle of socially desirable inputs. Research demonstrates that process-based fairness rooted in social psychology is a promising approach.[17] Procedural justice requires affording the community a voice and opportunities for participation, the use of neutral procedures for decision-making, treatment with respect and dignity, and communication of trustworthy motives through consideration of and responsiveness to people's needs and concerns. Decision-makers and community members can generate prosocial behavior over the long term by adhering to the principles of procedural justice. In effect, there are two goals for our prosocial production system: to limit negative experiences and to promote positive behavior.

A key factor in achieving prosocial engagement online, though, is alignment with platform business models. The business models are heavily metric- and product-driven.[18] One reason why platforms focus on identifying bad behavior and then demonstrating a particular consequential response is that those actions are easy to

platform issues a detailed set of rules that leaves little opportunity for the community of users to come up with their own rules. BEN BRADFORD ET AL., JUSTICE COLLABORATORY AT YALE L. SCH., REPORT OF THE FACEBOOK DATA TRANSPARENCY ADVISORY GROUP 31 (2019).

[17] Tyler and Meares have illustrated this in Tracey L. Meares & Tom R. Tyler, *Justice Sotomayor and the Jurisprudence of Procedural Justice*, 123 YALE L.J.F. 525 (2014).

[18] As stated by Venkatesh, many tech companies have adopted the myth of "product is governance." Working based on this myth, companies have deemed reliance on self-regulation by the users as inefficient for governance because of the volume and scale of content that needs to be governed. Sudhir Venkatesh, *The Myth of Platform Governance: How Product Culture Shapes Content Moderation in Technology Firms*, YALE J. L. & TECH. (Forthcoming 2021).

measure—not unlike arrests in the real world.[19] Those who govern the behavior of online communities (mainly platforms' employees and policymakers who are engaged with user and content moderation) will want to determine to what extent procedural justice has in fact helped achieve prosocial goals. Thus, another contribution of this paper is to provide potential benchmarks for measuring the success we think our theoretical approach can achieve. We offer an explanation for how platforms could create an experiment to measure any prosocial approach to community vitality.[20]

In summary, then, the overall approach can be depicted in the following table:

GOALS		**ACTIVITIES**		
		*On the **platform***		*In the **community***
Limit negative experience	→	Content moderation to lessen the amount of hate speech, false content, etc.	→	Lessen the impact of negative and false content on real world activities such as elections
Promote positive behavior	→	Encourage constructive interactions	→	Increase the resilience and vibrancy of real-world communities

[19] The report of the Facebook Data Transparency Working Group shows this similarity by drawing an analogy between "Facebook's prevalence measurement" and "commonly used measures of crime." BRADFORD ET AL., *supra* note 16, at 18-19.

[20] Platforms do commonly survey their users to test for what is often called "customer satisfaction" with the platform experience (user experience or UX) or feelings about their experience. These studies are typically not considered within a framework of the site's capacity to create positive social experiences with an eye to community building. The performance metric matters because time on the site is a profit indicator. In a business sense, time on a platform spewing hate speech and time promoting tolerance are both related to platform business model success.

Our argument is expanded below. Section II discusses how platform governance has evolved and led to the adoption of deterrence-based approaches. Section III theorizes a prosocial system that platforms and online communities could use to limit antisocial behavior, promote prosocial behavior, and potentially measure the effect of the prosocial system on their platforms. Section IV concludes the paper.

GOVERNING ONLINE BEHAVIOR

Our argument depends on an analogy to how the criminal legal system effectively and fairly addresses criminal behavior in the real world. In section A, we discuss the gradual change from community- to authority-based governance on platforms. Section B lays out the shortcomings of deterrence-based governance approaches on platforms, which are often authority-based, by drawing an analogy between such methods and criminal justice methods.

Emergence of Online Norms and Governance of Platforms

It is easy to imagine that developing and conforming to online terms of service is a straightforward matter, but in fact, online behaviors and the norms governing them developed gradually and under the influence of multiple social and legal pressures. It is useful to divide the periods of online platform governance into different phases. The first is one of community governance, which came about in the early 1990s, before the Internet became the overwhelmingly dominant communication mechanism. During this time, online communities with a wide range of interests and goals sprung up, but they relied on different technologies with their own social affordances,[21] and volunteers and system administrators ran various virtual spaces.[22] These online communities used the Internet and other communications technologies to achieve their goals,[23] which ranged from discussing their favorite shows, politics, and literature to organizing political gatherings, holding community meetings, and solving each neighborhoods' problems.

The Internet as a whole is broadly decentralized, and these early Internet-based communities usually governed themselves in a decentralized manner. Prior to the emergence of the web, there were various ways in which people gathered online in communities. Two prominent ones were Listservs (sometimes called "mailing lists" or just "lists," terms still sometimes in use today) and Usenet newsgroups (Usenet started outside the Internet and was in wide use for several years but has mostly ceased to play a role in people's

[21] Barry Wellman et al., *The Social Affordances of the Internet for Networked Individualism*, 8 J. COMPUT.-MEDIATED COMM. (2003); Laura W. Black et al., *Self-Governance through Group Discussion in Wikipedia: Measuring Deliberation in Online Groups*, 42 SMALL GROUP RES. 595 (2011).

[22] These distributed systems were run by system administrators, who managed their technical maintenance needs. As these communities grew, their system administrators did not want to get involved with governance and so asked the communities themselves to take part in decision-making. For example, platforms like LambdaMOO and Habitat made major changes to their governance and used community governance mechanisms such as "grassroots petitions" and "collective voting." Sherry Turkle, *Virtuality and Its Discontents: Searching for Community in Cyberspace*, *in* THE WIRED HOMESTEAD: AN MIT PRESS SOURCEBOOK ON THE INTERNET AND THE FAMILY 385 (Joseph Turow & Andrea L. Kavanaugh eds., 1996). LambdaMOO and Habitat were early online multi-user environments where people interacted with each other through pre-web technologies. Diane J. Schiano, *Lessons from LambdaMOO: A Social, Text-Based Virtual Environment*, 8 PRESENCE: TELEOPERATORS & VIRTUAL ENV'T 127 (1999); Chip Morningstar & F. Randall Farmer, *The Lessons of Lucasfilm's Habitat*, 1 J. VIRTUAL WORLDS RES. (2008).

[23] Black et al., *supra* note 21; HOWARD RHEINGOLD, TOOLS FOR THOUGHT: THE HISTORY AND FUTURE OF MIND-EXPANDING TECHNOLOGY (2000); MARC A. SMITH & PETER KOLLOCK, COMMUNITIES IN CYBERSPACE § 1 (1999); Constance Elise Porter, *A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research*, 10 J. COMPUT.-MEDIATED COMM. (2004).

online experience).[24] The nature of the technical operation of those technologies meant that multiple Internet site operators had some role to play.[25] It was not possible to take control of the operation of a site (a newsgroup, for example). These forums, mostly based on Listservs and Usenet newsgroups but sometimes on the early web, were distributed and decentralized in operation (even if, as in some cases, they depended on centralizing technology like the web). The systems were for the most part technically basic, so they depended on cooperative administration. Attempts to impose central control resulted in people objecting by setting up alternatives.[26]

The second phase started in the late 1990s, when the web became very popular. The Internet and the World Wide Web (often just called "the web") are not the same technology, and the difference may influence the governance models that emerge in each system. The Internet is a global network made up of many independent, globally interconnected networks. As online platforms grew, the networks specialized more or started using the more centralized technology of the web. However, early adopters of the web (such as Wikipedia and Slashdot) used "community

[24] Usenet was a global bulletin board that allowed user-to-user interaction through their local news servers. Users would send messages from their server to other users' servers, and they could communicate and react to each message. It is important to note that the web was technically distributed, but it also allowed for centralized governance. For example, it could turn the website operator into an exclusive intermediary (a service provider) because the operator could disallow user-to-user interaction, and the users had to communicate through the website. An example can clarify this: if Facebook removes a group from its website, the members no longer have access to that group under any circumstances. But if a server no longer hosts a newsgroup, the users could move to another server and have access to the same newsgroup. Bryan Pfaffenberger, *A Standing Wave in the Web of Our Communications*: *Usenet and the Socio-Technical Construction of Cyberspace Values*, *in* FROM USENET TO COWEBS 20 (Christopher Lueg & Danyel Fisher eds., 2003).

[25] PETER H. SALUS, CASTING THE NET: FROM ARPANET TO INTERNET AND BEYOND (1995).

[26] *Id*. at 144.

governance" mechanisms akin to those of older systems.[27] Only gradually did social media platforms adopt a hybrid authority-community governance or a more hierarchical, authority-based governance.[28] Before the emergence of centralized platforms, the typical virtual space was like a main street where communities grew. As platforms used technology that tended to encourage centralized operation and control, virtual spaces became more like shopping malls, and users turned into customers.[29]

The third phase started in the mid-2000s. Scholars warned that when the commercial stakes in online communities rose, so too would the interest in directing the participants' attention or controlling the format of interaction to suit the profit-making agendas of corporate partners.[30] It was around 2006 that the commercial stakes became high once certain platforms began amassing users and generating revenue by using their online platforms to regulate user behavior, rather than just facilitating communication. Some became multisided online markets, providing services other than facilitating communication. Economically, it was in these platforms' interest to keep users inside their "ecosystems." Examples of this pattern include some of the most familiar names in online platforms, such as Facebook, Twitter, or

[27] See the following articles for a more detailed account of Wikipedia and Slashdot governance mechanism: Cliff Lampe & Paul Resnick, *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space*, *in* PROCEEDINGS OF THE SIGCHI CONFERECE ON HUMAN FACTORS IN COMPUTING SYSTEMS 543 (2004); Aleksi Aaltonen & Giovan Francesco Lanzara, *Building Governance Capability in Online Social Production: Insights from Wikipedia*, 36 ORG. STUD. 1649 (2015); Laura Stein, *Policy and Participation on Social Media: The Cases of YouTube, Facebook, and Wikipedia*, 6 COMM. CULTURE & CRITIQUE 1 (2013).

[28] BRADFORD ET AL., *supra* note 16.

[29] Turkle, *supra* note 22.

[30] BRADFORD ET AL., *supra* note 16, at 30–31; Dara Byrne, *The Future of (the) 'Race': Identity, Discourse, and the Rise of Computer-mediated Public Spheres*, *in* LEARNING RACE AND ETHNICITY, YOUTH AND DIGITAL MEDIA (Anna Everett ed., 2008). Byrne explained that "As the commercial stakes in online communities rise, so too will the interest in directing the attention of participants, or controlling the format of interaction, to suit the profitmaking agendas of corporate partners."

Sina Weibo, but an exhaustive list is now impractical because the pattern is so widespread. The incentive to keep users in their "ecosystem" meant that, unlike pre-web systems, these newer platforms developed technical features that were only available inside that platform, so alternatives were not possible.

In platforms with authority-based governance, users went from being members of a particular online community to being subjects of the platform. This change in governance structure might have been because, as platforms' networks became larger, they did not think it feasible to leave their communities to govern themselves.[31] But it also might have been because the platforms' interests were better served by their power over their users.

This is not to say that "community governance" does not exist on online platforms anymore. Online platforms might adopt hybrid governance mechanisms that use several mechanisms for governing the behavior of users:

1. A top-down user agreement and a content moderation policy drafted by the platform's lawyers;
2. Community rules that the community generates within its various sub-groups;
3. Overall community rules (Netiquette, Reddiquette, or the like) which are not binding, but to which the community as a whole contributes and offers amendments.

Some might argue that Facebook and other centralized platforms are investing in features and policies that can empower their communities.[32] This is true, yet they still have dominant

[31] BRADFORD ET AL., *supra* note 16, at 30.
[32] For example, Facebook creates a sense of community by "group building." It states that "Facebook gives you powerful tools to help your group thrive. These

authority-based governance in place. For example, Facebook empowers its community to convene various groups and set up their own rules and code of conduct. But the community (members of the group) does not have much say in policy changes. Facebook has well-elaborated community standards (mainly drafted by Facebook lawyers) that impose restrictions on many aspects of individuals' and communities' behavior. It does not leave much room for self-governance.[33]

A better example of the hybrid model is Reddit, which has its own terms and conditions and imposes standards of behavior but also allows communities of users to assert their own rules. Reddit emphasizes the community and the role of the moderators, explaining that it rarely wants to get involved with content moderation: "Reddit may, at its discretion, intervene to take control of a community when it believes it in the best interest of the community or the website. This should happen rarely (e.g., a top moderator abandons a thriving community), but when it does, our goal is to keep the platform alive and vibrant, as well as to ensure your community can reach people interested in that community."[34] Hence Reddit does intervene, but the basic interaction is first within

focused tutorials give you more information on these helpful features and how to use them." *Using Key Group Tools*, FACEBOOK, https://www.facebook.com/community/using-key-groups-tools/ (last visited Jan. 24, 2021).

[33] BRADFORD ET AL., *supra* note 16, at 31. In a study about Facebook, YouTube, and Wikipedia, Stein explained that users did not know about Facebook policy changes until they came into effect, and they challenged Facebook's policies about issues such as privacy. She concluded that (at the time of writing the paper) Facebook and YouTube gave their users minimal control over content and governance of the website. Stein, *supra* note 27, at 354. It is of note that Facebook undertakes meetings with third-party experts to discuss its policies and might make changes accordingly, but third-party experts might not be community members. One of the first consultations of this kind happened in 2012 at Stanford. See Klonick, *supra* note 13.

[34] *Moderator Guidelines for Healthy Communities*, REDDIT, https://www.redditinc.com/policies/moderator-guidelines (last visited Jan. 20, 2021).

the community, and the platform rules are secondary. Reddit also has an "informal expression of value"[35] called Reddiquette that communities refer to, in addition to the formal terms and conditions provided by Reddit.[36] Because Reddiquette is a normative system based on Reddit users' values, it is more acceptable to users than top-down platform rules.[37] So, Reddit offers examples of all three modes of the hybrid governance mechanism at once.

Another example of a hybrid governance mechanism is Wikipedia. Wikipedia's policies mainly come from its community of editors. Similar to Reddit's Reddiquette, English-language Wikipedia has a Wikiquette. Community editors have written down behavioral standards that can be changed by community members through consensus.[38] However, Wikipedia also has a top-down mechanism that involves its legal team. That mechanism can be invoked to make decisions that overrule the community. For example, their trust and safety group can govern its users' behavior. The group's Wiki page indicates that it "aims to defer to local and global community processes to govern on-wiki interactions." While acknowledging that intervention may happen rarely, it also states that they step in to protect the safety and integrity of users, contributors, and the public.[39]

[35] *Reddiquette,* REDDIT (2020), https://reddit.zendesk.com/hc/en-us/articles/205926439-Reddiquette.

[36] *Content Policy*, REDDIT, https://www.redditinc.com/policies/content-policy (last visited Jan. 20, 2021).

[37] Fiesler et al. illustrated this by undertaking empirical research and concluded, "It is also much more common for subreddits to refer to Reddiquette than official policy, suggesting again that the rules closest to the community itself are the most visible, prioritizing an individual subreddit over Reddiquette over Reddit policy." Casey Fiesler et al., *Reddit Rules! Characterizing an Ecosystem of Governance*, *in* 12TH INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA 78 (2018).

[38] *Etiquette,* WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:Etiquette (last visited Jan. 20, 2021).

[39] *Trust and Safety,* WIKIPEDIA, https://meta.wikimedia.org/wiki/Trust_and_Safety (last visited Jan. 20, 2021).

Repeating the Same Mistake Online

The criminal legal system relies heavily on sanctions or threat of them to achieve behavioral governance goals. The three-strikes laws and mandatory minimum sentences and guidelines to increase penalties for certain crimes are but two examples of applying mandatory sanctions to certain behaviors.[40] As we have explained, this approach to governing people's behavior centers deterrence as a theory of compliance. Deterrence-based approaches focus on increasing the cost of rule-breaking so that people, out of self-interest and fear of punishment, do not break the rules. These approaches have notable weaknesses, however. They are most effective in situations where surveillance is possible, and because they depend so much on surveillance, in the real world they can be extremely costly.[41]

In the realm of social media, governance mechanisms and the popular rules that platforms usually implement also rely on deterrence. These systems focus mainly on compliance and individual violations. Their common punitive measures are analogous to those in the criminal justice system. Platforms typically suspend accounts in the face of infractions—the functional equivalent to putting someone in "jail," which can operate to incapacitate a person or punish them or both. Sometimes, after multiple violations, a platform might ban a user from the platform entirely (analogous to so-called "three strikes" laws).[42]

[40] Andrew V. Papachristos et al., *Why Do Criminals Obey the Law? The Influence of Legitimacy and Social Networks on Active Gun Offenders*, J. CRIM. L. & CRIMINOLOGY 401 (2012).

[41] TYLER, *supra* note 7, at 263; Tracey Meares, *Broken Windows, Neighborhoods, and the Legitimacy of Law Enforcement or Why I Fell in and out of Love with Zimbardo*, 52 J. RES. CRIME & DELINQ. 609 (2015).

[42] Scholars have drawn an analogy between mechanisms of the criminal justice systems and these platforms' governance approaches in the past. This might especially be because the line between digital and real life has been blurred, and

Content moderation as it operates today is similar to focusing on "arrest rates" and "crime rates" in the criminal legal system.[43] Both mechanisms are outcome-oriented and do not have as goals either prevention or reform. Rather than attempting to change users' behavior through education or even mere notice of the rules, the major focus of many platform moderation efforts is simply to count and reduce individual violations. These "elimination" measures do not positively contribute to users' behavior; for example, they do not encourage users not to repeat the offense. However, Tyler et al. undertook an experimental study about Facebook which empirically showed that the users that Facebook treated fairly during content moderation were more likely not to repeat the offense than those who were not treated fairly.[44]

There is a link between the growth and change of structures for online interaction and the mode of governance. As the Internet grew and main streets turned into shopping malls, platforms increased their use of aggressive methods like content takedowns and blocking and suspending accounts. We contend that in shifting from communal governance approaches to more authority-based ones, platforms started making the same mistakes that the judicial system and the police now make: focusing on individuals and

users' lack of access to these platforms might highly affect their access to their community, families, and friends. It can even give them the feeling that their access to their community was cut off as a result of a platform's suspension or ban. Tyler et al. argued that account suspension or cancellation "parallels" some criminal justice mechanisms, such as incarceration. Tyler et al., *supra* note 2. Other scholars have also mentioned that platforms' governance approaches are punitive, and they tend to adopt or are more likely to use methods similar to criminal justice. Platforms' inclination toward punitive governance approaches is elaborated in the following sources: Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 New Media & Soc'y 4376 (2018); Sarita Schoenebeck et al., *Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment*, 5 PACM ON HUM.-COMPUT. INTERACTION (2021).

[43] BRADFORD ET AL., *supra* note 16, at 19.

[44] Tyler et al., *supra* note 2.

eliminating or weakening the communities with authority-based governance. This is of course not to say that these efforts have so far led to the pernicious effects of what we see in the criminal justice system. Importantly, the web itself is only thirty years old, and online experience is still but a fraction of human life. Our point is simply this: It is unwise to continue to build models of online governance founded on assumptions similar to those that have had poor effects in criminal justice. Since better approaches for criminal justice have already been proposed, perhaps those models can also be applied effectively in the online world.

IMAGINING PROSOCIAL PRODUCTION ON PLATFORMS

So what does a good alternative governance approach on platforms look like? In this part, we argue that focusing on community vitality as a goal rather than merely identifying and punishing bad behavior motivates the production of prosocial governance mechanisms that can facilitate compliance, engagement, and cooperation. Relying on the theory of procedural justice, we conceptualize a prosocial production system that could lessen violations on platforms by motivating individuals to internalize rules, voluntarily comply with them, and engage in healthful interaction.

Before focusing on community vitality and applying procedural justice, we cover several important work that scholars and practitioners have done in this space. Eli Pariser, a tech-entrepreneur, has undertaken several initiatives that focus on communities on social media platforms. For example, Civic Signal (new public) works on creating “public spaces” on platforms.[45] They also work on creating vibrant, livable online spaces. One of the

[45] NEW PUBLIC, https://newpublic.org/ (last visited June 14, 2021).

inspiration for this work is Jane Jacobs, an urbanist and activist who objected to the elimination of communities and social structure. She also fought against building highways and fake parks in the suburbs at the expense of demolishing communities.[46] Drawing an analogy between Jacob's work on neighborhood and communities, digital activists and engineers have tried to envision social structures on platforms.[47]

As we mentioned, procedural justice in decision-making is central to creating vital communities. Research demonstrates that four factors matter.[48] The first is participation or voice: the decision-maker should give people the opportunity to explain their situation and perspective.[49] Participation in decision-making processes should happen at various stages. This means that people's voices should not exclusively be heard after a dispute arises, but they should be able to take part in different stages of decision-making processes that can affect them. This can be during any or all of policy-making, dispute resolution, or enforcement processes. Second, people care about being able to ascertain whether authorities are being fair as they carry out decisions. Fairness includes the following: neutrality, objectivity, factuality of decision-making, consistency in decision-making, and transparency.[50] Third, people care a great deal about being treated with dignity and respect. People care about how their community leaders and authorities treat

[46] JANE JACOBS, THE DEATH AND LIFE OF GREAT AMERICAN CITIES (2016).
[47] Amy X. Zhang et al., *PolicyKit: Building Governance in Online Communities*, *in* PROCEEDINGS OF THE 33RD ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY 365 (2020). Zhang et al have come up with a software design that focuses on designs that could potentially lead to vital communities and plurality in governance.
[48] *Procedural Justice*, *supra* note 9, at 103; BRADFORD ET AL., *supra* note 16.
[49] POLICE LEGITIMACY, supra note 9.
[50] Tom R. Tyler & Cheryl J. Wakslak, *Profiling and Police Legitimacy: Procedural Justice, Attributions of Motive, and Acceptance of Police Authority*, 42 CRIMINOLOGY 253 (2004).

them; they usually respond positively to being treated with dignity, respect for their rights, and politeness.[51] Finally, people want their leaders and decision-makers to act out of a sense of benevolence toward them, so it is important that they perceive authorities to be communicating trustworthy motives. People attempt to discern why authorities act the way they do, and a procedurally just decision-making process gives them the perception that the authorities are benevolent, well-intentioned, and sincere, and do not act only out of self-interest.[52]

Procedural justice is central to the creation of a self-motivated prosocial production system—that is, a system that by its nature produces a cycle of socially desirable inputs. In the following sections, we argue that platforms can motivate voluntary rule-following through procedural justice. Importantly, members of online communities can cooperate with each other and the authorities to lessen the impact of rule violations and negative behavior. We also argue that this approach can do more than motivate rule-following. We think prosocial approaches are an additional step platforms can take to encourage their communities to actively do good.

Prosocial Compliance and Cooperation: Limiting Antisocial Behavior on Platforms

The traditional goal of content regulation is to avoid harm by limiting negative content that violates platform rules. Prosocial approaches begin with the goal of limiting negative content but are also concerned with the objective of promoting positive content. Prosocial approaches treat a positive social environment as a goal.

[51] *Id*. at 253.
[52] Tom R. Tyler & E. Allan Lind, *A Relational Model of Authority in Groups*, 25 ADVANCES EXPERIMENTAL SOC. PSYCHOL. 115 (1992).

The goal of achieving a positive environment has two aspects. The first is to aid traditional regulation. When platforms enforce their rules through traditional control mechanisms, they identify and sanction undesirable content. This motivates users to evade platform authorities and hide their actions. However, when users identify with and feel positively about the provider and their online community, they become more self-regulatory. To put it simply, they are more likely to want to do the right thing and to do it voluntarily. Hence, building a positive online climate facilitates effective regulation. An important part of this positive climate is accepting the legitimacy of the platform, its rules, and its enforcement mechanisms. When legitimacy is high, the threat of sanctions in not the primary means of promoting rule adherence.

The second goal is for the platform to serve as a safe space within which the members of different communities can interact constructively and civilly to address their common issues and concerns. This positive climate will make the time that people spend on a platform more satisfying and will also enhance the possibility of useful dialogue about potentially divisive issues. That dialogue can then spill over into real-world communities and enable them to jointly address their political, social, and economic issues. Again, the legitimacy of the platform as an "honest broker" seeking to create a secure and safe space for such discussions is crucial. A key antecedent of legitimacy is the belief that an authority is benevolent and sincere, seeking in good faith to help people define and address their needs and concerns.[53]

[53] See Tom Tyler, *Policing in Black and White: Ethnic Group Differences in Trust and Confidence in the Police,* 8 POLICE Q. (2005).

Tyler defines legitimacy as the belief that authorities have the right to dictate proper behavior.[54] Meares defines legitimacy as a collection of individuals' perceptions of the laws and the authorities that enforce them.[55] People comply with the law as long as they perceive the authorities and their laws as legitimate. When people perceive authorities as legitimate, they will largely regulate their own behavior, so hierarchical enforcement mechanisms will be less necessary.[56] By resorting to procedural justice (instead of deterrence-based mechanisms) and building the legitimacy of decision-makers, it is possible to encourage people to comply with the rules.

To limit antisocial behavior on platforms, we need to go beyond compliance and address cooperation. Cooperation includes the willingness to accept authority, deference to the decisions made by the authority, and everyday rule adherence. Cooperation is also the willingness to aid decision-makers (the authorities in a governance mechanism) in identifying violations and wrongdoers and helping with the adjudication of conflicts. Tyler and Jackson demonstrated that cooperation can be achieved by trust and confidence in an authority; it can also be achieved by normative alignment, or sharing the authorities' goals and values.[57]

A crucial question is, how can platforms generate compliance and cooperation? Recall that procedural justice suggests that people are more likely to follow rules when they participate in

[54] Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283 (2003).

[55] Tracey L. Meares, *Norms, Legitimacy and Law Enforcement*, 79 OR. L. REV. 399 (2000).

[56] Tom Tyler & Steven Blader, *Can Businesses Effectively Regulate Employee Conduct? The Antecedents of Rule Following in Work Settings*, 48 ACAD. MGMT. J. 1143 (2005).

[57] Tom R. Tyler & Jonathan Jackson, *Popular Legitimacy and the Exercise of Legal Authority: Motivating Compliance, Cooperation, and Engagement*, 20 PSYCH. PUB. POL'Y & L. 78 (2014).

the decision-making process and feel that the decision-maker has heard their voice. Translating this participation into the online environment can take various forms. For example, community groups might come up with their own rules, or community members might get meaningful participation in the policy-making process, or community members might receive fair treatment during the adjudication process. Establishing a platform's legitimacy might be harder than it is for other authorities that have been approved by their community members and are appointed through a democratic process. Platforms' decision-makers that adjudicate disputes and enforce rules are the employees of the platform and not selected or appointed through a democratic process. This might affect users' incentives to follow rules, since they might not buy into the outcome of the adjudication, so they might try to subvert the norms and the outcomes that the platform tries to enforce.

When authorities over a social group treat group members fairly, the members feel included, find the group valuable and valid, and identify with its values.[58] The fair treatment preference is constant in various settings and different communities. Diversity in ethnicity, location, and other aspects does not usually affect individuals' preference for fair treatment.[59] This is especially important in the context of platforms because they serve global and diverse communities. This is not to say that fairness of their

[58] WHY COOPERATE, *supra* note 9.

[59] Tyler and Huo examined whether race or ethnicity has an impact on authorities' personal perception that can weaken procedural justice generality. TOM R. TYLER & YUEN J. HUO, TRUST IN THE LAW 153 (2002). Other scholars have tested the generality of procedural justice and its applicability to various settings. So far, a couple of studies have come to the conclusion that procedural justice leads to cooperation and compliance across different settings. Scott E. Wolfe et al., *Is the Effect of Procedural Justice on Police Legitimacy Invariant? Testing the Generality of Procedural Justice and Competing Antecedents of Legitimacy*, 32 J. QUANTITATIVE CRIMINOLOGY 278 (2016); JONATHAN JACKSON ET AL., JUST AUTHORITY? TRUST IN THE POLICE IN ENGLAND AND WALES (2012).

treatment has an absolute effect on people. There are circumstances under which fairness might not motivate cooperation.

Fair treatment requires decision-makers to be objective and neutral. On social media platforms, we can detect fairness or unfairness during the enforcement process. However, to go beyond applying procedural justice to dispute resolution and enforcement processes, it is important to consider the fairness of interactions between platform decision-makers and platform users as well as among community members. Fairness of interactions (for example, showing tolerance to opposing views and considering all arguments based on merit) can lead to building a community with members that perceive the processes and decision-makers to be fair, which leads to further cooperation within the community.

Offline or online, people care about being treated with dignity and respect during interactions, whether with other community members or with the decision-makers on the platform.[60] Elimination of disrespectful content does not in itself afford people such respect. However, an increase in respectful treatment can provide people with what they desire and also provide a chance to cooperate with authorities.

Communicating trustworthy motives might be especially important in the case of unelected authorities, whether they are platform owners or community leaders who are not elected by the community members. Commercially driven initiatives and their commercially driven authorities, especially, should make sure not to communicate only profit-making incentives when they make and enforce decisions that affect the community. To be effective, they should have the best interests of the community in mind, avoid

[60] In our paper about Facebook, we demonstrated that users in an online setting care about procedural fairness. Tom Tyler et al., *supra* note 2.

acting merely out of self-interest, and communicate all of that effectively.

We can see elements of participation especially on some platforms with hybrid governance models, since they allow their users to participate in decision-making processes. For example, Nextdoor (a neighborhood social media platform) allows the community member volunteers, neighborhood leads, and group admins to make decisions and enforce Nextdoor's guidelines.[61] The users also get to vote about whether to remove a given piece of content. When the votes pass a certain threshold, the lead for the neighborhood takes the content down.[62] This is a good way to get people to cooperate with the authorities of groups—in Nextdoor's case, the leads of the neighborhood.

To some extent, it is possible to compensate for the shortcomings of top-down rules by being procedurally just. As Tyler et al. showed in their paper on Facebook's governance model, when rule violators on Facebook were treated with procedurally just adjudication mechanisms, they were less likely to repeat the violations. Therefore, as Tyler et al. concluded, the users were more likely to self-regulate and follow the top-down rules when Facebook exercised procedural justice in its dispute resolution process. These findings speak to the first issue noted: the desirability of self-regulation in response to viewing the platform as a legitimate authority.

A further goal not addressed in the Facebook study is the ability of these same fair procedures to enhance the online climate on a site. We will discuss this goal in the next section in more detail.

[61] *About Moderation,* NEXTDOOR, https://help.nextdoor.com/s/article/About-moderation? (last visited Jan. 23, 2021).

[62] *About Community Reviewers and Moderation,* NEXTDOOR, https://help.nextdoor.com/s/article/Community-Reviewers-and-Moderation?.

The goal of many platforms is to create a safe climate within which people can constructively discuss emotional and potentially divisive issues in their lives and communities. People's ability to do so is also affected by whether they trust the authorities creating and managing the platform through which they are interacting. Again, legitimacy is key to providing a baseline level of comfort and reassurance that can enable such dialogue.

Promoting Prosocial Behavior

Prosocial approaches are different from simply trying to avoid harms or violations, no matter whether one is focused on criminal justice system outcomes or trying to ensure compliance with content moderation rules online. Prosocial approaches treat a positive social environment as a goal.[63] If the platform and its users create a positive social environment, the need for control by the platform is reduced because that social environment produces socially desirable outcomes.

While procedural justice—based approaches can enhance rule-following by motivating voluntary compliance, we think prosocial approaches based on procedural justice theories can do more: platforms can use them to motivate users to do good. To promote prosocial behavior, platforms must increase community engagement, individuals' desire to pursue a collective goal, and engage in economic and political activities. Engagement is involvement with one's own community.[64] Specifically, it is discretionary cooperation, meaning that instead of just following the rules authorities impose and cooperating with the authority, the community proactively behaves in such a way that the members

[63] For a deeper understanding of prosocial approaches in criminal justice system and online platforms, refer to the website of Justice Collaboratory based in Yale Law School: https://law.yale.edu/justice-collaboratory.
[64] Tyler & Jackson, *supra* note 57.

trust one another and know that if a problem arises, they can face it collectively.[65]

For a community to engage (offline or online), the individuals must identify with the values of the community and be willing to act on behalf of the collective. The decision-makers and authorities can incentivize the community to engage with one another by being legitimate. Engagement can increase when the community members have normative alignment with one another and identify with the values and goals of the community.[66]

One approach to increasing engagement is to create virtual social structures. According to prosocial theories, social structures can create opportunities for communities to thrive and cooperate.[67] These structures in real life are gyms, town halls, youth centers, bars, bistros, and the like. These social structures are the heart of community vitality in the real world.[68]

We can translate social structures to their online analogues. For example, online forums, town halls, and groups, and even some algorithms and other virtual tools, can play a role in building a strong social structure. Black et al. also mentioned that even simpler communication systems such as email lists can help in providing social structures. Byrne argued that the virtual "forums" on websites are where community vitality is happening and people engage. He called these forums central to public life and an opportunity to

[65] *Id*. at 81.

[66] *Id*. at 84.

[67] As Meares argued, where social structures are weak, it is difficult to exert social control. Thus, to be able to govern online communities through self-regulation and social control, it is necessary to provide the social structures. Meares, *supra* note 55.

[68] RAY OLDENBURG, THE GREAT GOOD PLACE: CAFES, COFFEE SHOPS, BOOKSTORES, BARS, HAIR SALONS, AND OTHER HANGOUTS AT THE HEART OF A COMMUNITY (1999).

understand how various communities construct, modify, and stabilize.[69]

Historically, community vitality was generated through instant messengers, chat rooms, weblogs, and discussion boards. For example, chat rooms became the social structures where users could discuss the rules and responsibilities governing their behavior in their online community. The effect of cyberspace on physical world communities, not to mention the fact of online communities that depend on cyberspace for existence, has been profound. The effect even inspired predictions that as bars, restaurants, and other places came to lose their sense of community vitality, perhaps online communities would replace them and bring community vitality.[70]

To advance prosocial interaction, platforms must enhance engagement with political, social, and economic activities. Tyler and Jackson identified the following indicators of engagement (actions to help the community and its vitality):[71]

- Perceived social capital (community members helping each other and working together to bring safety)
- Community identification (being proud of your community)
- Political capital (engaging with changing political decisions)
- Economic activities (going to shops and restaurants and spending time with the community)

Tyler and Jackson argued further that procedural justice is associated with indicators of engagement. We can theorize that if procedural justice criteria are satisfied in an online group, it is likely that engagement also will increase. The theory's hypothesis is that

[69] Byrne, *supra* note 30.
[70] OLDENBURG, *supra* note 68.
[71] Tyler & Jackson, *supra* note 57, at 79.

if people are treated well (fairly, with respect and dignity) by those they encounter in a given community, they are more likely to engage with voluntary actions, build social capital, and get involved with economic activities. In the next section, we describe how to measure a prosocial production system. That way, the measures can feed back to generate the desired behavior and meet the criteria stated above.

Improved Measurements for a Prosocial Production System

How should platforms create an environment in which prosocial activity begets more prosocial activity, creating a positive feedback loop that ensures a good online social environment? In other words, how should platforms set up a prosocial production system? The first step for the platforms is to select a prosocial goal for themselves. The goal could be to achieve healthy interaction or enhance civility. In order to operationalize "healthy interaction" or "civility," we define them and determine the constitutive elements. For example, we can operationalize civility by asking the extent to which a candidate action exhibits tolerance and respect. It is also important to have an understanding of what constitutes tolerance and respect and how to measure the increase or decrease of each. Using legal and social science methods, we can discover the constitutive elements of respect and tolerance.[72]

[72] Scholars across various disciplines have discussed how to define and operationalize prosocial goals such as civility and healthy online interactions. *See* Jeremy Waldron, *Civility and Formality*, NYU SCHOOL OF LAW, PUBLIC LAW RESEARCH PAPER NO. 13-57 (2013); Zizi Papacharissi, *The Virtual Sphere: The Internet as a Public Sphere*, 4 NEW MEDIA & SOC'Y 9 (2002); Arthur Santana, *Virtuous or Vitriolic: The Effect of Anonymity On Civility in Online Newspaper Reader Comment Boards*, 8 JOURNALISM PRACTICE 18 (2014); Myiah Hutchens et al., *What's in a Username? Civility, Group Identification, and Norms*, 16 J. INFO. TECH. & POL. 203 (2019). Chris Vargo & Toby Hopp, *Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter: A Congressional District-Level Analysis,* 35 SOC. SCI. COMPUT. REV. 10 (2017).

The next step is to set up the virtual social structures for a sample of individuals. For example, as discussions are heating up and are becoming controversial on some general thread, the platform can empower the poster by recommending the creation of a group. There must also be policies and methods that encourage people to do good—for example, prompts that would pop up in the form of pithy messages when users join a group are having a conversation.

Finally, it is critical to measure these efforts. Platforms are run (perhaps even "overrun") with attention to metrics. If they cannot measure it, they will not do it. We have identified the need for "measurement" in our conversations with platforms when discussing strategies for enhancing healthful interactions. Using metrics and measurements is a good way to improve decision-making processes; however, the platforms need to enhance and modify their approach and update their metrics. Thus, in this paper, we also provide some suggestions and benchmarks for measuring social phenomena, with the hope to improve and standardize measurement benchmarks on platforms. [73]

In collaboration with one platform, we have undertaken a study that implements some of these suggestions. For example, we have designed prompts and messages based on the procedural justice indicators. These prompts are displayed when the users enter a virtual social structure such as a group. The prompts encourage and remind the community members of the platform's guidelines and ask the community members to "listen to each other" (allowing for participation), "lead with compassion" (respect others with

[73] Other scholars have also come up with methods to measure prosocial behavior online. For example, see Jiajun Bao et al., *Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations*, *in* PROCEEDINGS OF THE WEB CONFERENCE (2021). Bao et al. have created a process through which we can quantify prosocial outcomes on platforms.

dignity), “cite sources” (maintain neutrality and be objective), and “take other people’s issues seriously” (show good faith).

Measuring how much the prosocial governance mechanism produces prosocial behavior can help reform the governance mechanisms based on science and not clairvoyance. To measure prosocial compliance, which means feeling obligated and motivated to follow rules, we need a less outcome-oriented approach than the one usually followed by platforms. Prosocial compliance does not only mean that people comply with the rules, but also that people self-regulate and do not turn into repeat offenders.

One solution for a less outcome-oriented approach is to use community as the unit of analysis. Thus, instead of measuring only relations between or among individuals, we should measure individuals’ relations with the community. To measure whether people have internalized rule-following, as Tyler et al. have previously done,[74] we can use survey strategies to measure whether providing virtual social structures and treating the users with procedural justice has had any effect on following the rules. The surveys should also ask why the members followed the rules, out of self-interest or out of norm alignment with the community, and the perceived legitimacy of the decision-makers.

We can measure community engagement through the indicators mentioned in the previous section: do they volunteer to help their online community members, are they proud of the online community they belong to, have they accumulated social or political capital and engaged more with economic activities? Through a survey, people can indicate how likely they are to attend online political activities on the platform, get engaged with transactions, or

[74] Tyler et al., *supra* note 2.

intervene if they see members being disrespectful to each other, as well as whether they are proud or feel good about being involved with an online group.[75]

This method is, however, insufficient, as it measures the users' opinions *post-factum* and is not an observation of actual behavior. There are other methods that can measure prosocial behavior during ongoing interactions. For example, an indicator for cooperation is community-led efforts to inform the authorities of a problem. We can also control for the increase or decrease in the number of voluntary initiatives that community members come up with in order to help the decision-makers and the community leaders bring more civility to the platform or increase healthy interactions.

An additional way is to control for changes in prosocial indicators by observing the communities' social, political, and economic activities on the platform. For example, we can consider an increase or decrease in participation in voting and creating sub-communities to discuss politics. We can also measure the increase or decrease in participation in collective actions (for example, an online fundraising event). If the platform is multisided, i.e., it facilitates transactions as well as interactions, engagement can also be measured by controlling for an increase or decrease in economic activities (the rate of buying and selling on the platform).

A common approach to measuring the effect of governance mechanisms that communications and human—computer interaction scholars use is sentiment and textual analysis. Scholars use sentiment analysis and textual analysis to measure offensive

[75] We mentioned the criteria in *supra* Section II(B). As a reminder the criteria are: Community identification (being proud of your community), Perceived social capital (community members helping each other and working together to bring safety), Political capital (engaging with changing political decisions), and Economic activities (going to shops and restaurants and spending time with the community).

words or hate speech in a certain corpus of text (in the case of a platform, some set of messages on it).[76] The software often works based on a lexicon that can assess the tone of the text and label it as positive, negative, or neutral.[77] Software often has a training component so that the software can be tailored within some limits to the likely normal baseline of sentiment found in an average text from a given source. The trained software can measure the rate of positive, negative, and neutral words based on the number of occurrences and provide an estimate of how negative or positive certain texts are. The positive and negative sentiments can be correlated with various prosocial values—for example, the positive sentiment can be civil interactions, and the negative can be uncivil interactions. However, it is critical to first train the software with what is perceived as civil or uncivil to attain better results

CONCLUSION

Over time, as platforms became both commercialized and centralized, their approach to governance changed. Instead of fostering the communities that existed on their platform, they used a top-down deterrence-based mechanism to govern their platforms. This meant that they did not work on creating tools for empowering these communities, but tools to govern their users (primarily on an individual basis) and content. Compliance in these platforms does not mean motivating users to comply with the rules and regulations. Rather, platforms force their users to comply through deterrence-

[76] There are many studies that use this method, using different software and hate-speech or offensive speech lexicons. One example is Rishab Nithyanand et al., *Measuring Offensive Speech in Online Political Discourse*, *in* 7TH USENIX WORKSHOP ON FREE AND OPEN COMMUNICATIONS ON THE INTERNET (2017).

[77] Papacharissi, *supra* note 72; Federico Neri et al., *Sentiment Analysis on Social Media*, *in* Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining 919 (2012).

based mechanisms such as removal of content, suspension, and blocking.

We believe one way to reform the governance mechanisms of these platforms is by designing and implementing a prosocial production system. In this paper, we have presented a self-motivated prosocial production system to reform platforms’ punitive approach to governance, successfully limit negative behavior, and promote positive behavior. We have conceptualized the process through which we apply the theory of procedural justice to platforms’ governance mechanisms. We have also laid out the steps for designing a prosocial production system. Finally, we have presented a system through which various elements necessary for community vitality and prosocial behavior can be measured.

ONLINE REPUTATION SYSTEMS AND THE THINNING OF TRUST

Paolo Parigi & Dan Lainer-Vos[*]

INTRODUCTION

Trust, the skillful suspension of doubt, plays a crucial role in social life and online markets.[1] Two-sided marketplaces, in which two sets of agents exchange goods or services through an online intermediary platform, depend on trust cultivation among strangers.[2] eBay, Uber, Lyft,Airbnb, and other operators of two-sided market platforms rely on buyers and sellers trustingthe information each provides regarding the payment, quality, safety, performance of the advertised product or services.[3]

In the absence of product standardization or top-down sanctioning of defectors, reputation is the key mechanism that generates trust between a buyer and a seller.[4] The platforms that operate in online markets build sophisticated reputation systems to facilitate commerce. Onlinereputation systems consist of two parts: reviews and a set of ratings. Reputation systems differ in the prominence they give to one piece instead of the other and in how they display ratings.

[*] Paolo Parigi, Associate Director, Computational Social Science at IRiSS and Researcher, Stanford University & Facebook; Dan Lainer-Vos, Professor, Department of Sociology, USC.

[1] RACHEL BOTSMAN, WHAT'S MINE IS YOURS: THE RISE OF COLLABORATIVE CONSUMPTION (1st ed., 2010).

[2] Scholars use the term sharing economy or gig economy to describe this phenomenon. We reject the idealized understanding conveyed by the term "sharing economy," and since our focus is on exchange and trust relations and not on production, we opt to describe this new arena using the term "market" rather than economy.

[3] ARUN SUNDARARAJAN, THE SHARING ECONOMY: THE END OF EMPLOYMENT AND THE RISE OF CROWD-BASED CAPITALISM (2016).

[4] Bruno Abrahao et al., *Reputation Offsets TrustJudgments Based on Social Biases Among Airbnb Users*, 114 PROC. NAT'L ACAD. OF SCIS. 9848 (2017).

Nevertheless, reputation systems' ubiquity attests to their critical role in fostering trust in two- sided online markets.[5] Reputation facilitates trust between the two sides, and trust underpins these markets' functioning.

The development of two-sided markets and sophisticated reputation systems fundamentally alters the nature of trust. In traditional markets, trust is a cumulative byproduct of repeated dyadic exchanges. Trust is different from blind faith or purely calculated decision. It requires a skillful suspension of doubt on the basis of limited information. As partners engage in repeatedface-to-face transactions, mutual trust emerges. In these simple markets, trust was an interactional accomplishment. In such markets, inefficiencies in the diffusion of information constrain the exchange scale.

On the other hand, in two-sided online markets, face-to-face interactions are non-existent or minimal, and trust relations are technologically-mediated.[6] Furthermore, trust is not a cumulative result of repeated dyadic interactions in two-sided markets but an accretive product of the crowdsourced reviews generated by previous sellers and buyers. The technological mediation of two-sided markets enables rapid and efficient diffusion of information and fostersexchange relations at a previously unimaginable scale.[7]

The emergence of two-sided markets also alters the meaning of trust. We refer to this development as the *thinning* of trust. Whereas previously, sellers' trustworthiness took time to cultivate

[5] Paul Resnick & Richard Zeckhauser, *Trust among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System*, *in* THE ECONOMICS OF THE INTERNET AND E-COMMERCE (ADVANCES IN APPLIED MICROECONOMICS) (2002).

[6] MICHAEL MUNGER, TOMORROW 3.0: TRANSACTION COSTS AND THE SHARING ECONOMY (2017).

[7] SUNDARARAJAN, *supra* note 3.

because it depended on person-to-person interactions, now its growth is outsourced to the platform. Trust in traditional markets is mostly personal. It is associated with an individual partner to exchange, and it operates across domains. In contrast, trust in two-sided online platforms is impersonal because the judgment of trustworthiness is relative. We determine that seller X is trustworthy based on the aggregate assessment of others and relative to the judgment of many others on the reliability of other sellers of the same good or service.[8] As a consequence of the impersonal and technologically-driven nature of two-sided markets, the trust that underpins them is *thin* and confined to specific domains such as traveling or lodging.

While online reputation systems have made it easier to trust strangers and facilitate the circulation of trust, they have also removed part of the process we used to learn about one another. This disenchantment is reminiscent of Max Weber's argument on the rationalization ofreligion at the dawn of modernity.[9] Weber argued that Protestantism rationalized religion by eviscerating magic from religious practice and life. Similarly, we argue that the rise of two-sidedmarkets and the centrality of reputation systems had led to the disenchantment of trust.

Instead of trust being a spontaneous byproduct of interpersonal interaction, thin trust in two- sided online markets demands methodical cultivation,[10] is mostly impersonal, and is

[8] 93% of positive reviews for a vendor on Amazon or eBay translates to "trustworthy" only if other vendors on the
same page have similar or lower ratings.

[9] MAX WEBER, THE PROTESTANT ETHIC AND THE SPIRIT OF CAPITALISM WITH OTHER WRITINGS ON THE RISE OF THE WEST (2008).

[10] That thin trust requires methodical cultivation is evident in the proliferation of reputation management services that help vendors sustain their reputation in a volatile online environment.

domain- specific.

The thinning of trust provides a vivid illustration of the opportunities, risks, and responsibilitiesthat today's social sciences face. In the past, social scientists were, to the most part, observers of society, and their tools had little impact on social organization.[11] The migration of much of our social life to a digital interface, the fact that we shop online, for instance, creates an explosion of data, and more importantly, an exponential growth in our ability to intervene, manipulate, and curate social relations. A small change in the organization of a reputation system, for instance, can be extraordinarily consequential to the parties involved in the exchange. The capacity to measure and intervene is not limited to the domain of trust. Rather itimpacts many domains of social life from the most intimate (dating and romantic relationships) to the more collective and discrete (neighbor relations). The process that disenchanted trust is, in other words, repeating itself across all the domains that technology is making measurable.

Currently, platforms leave the design of the techniques that facilitate human interactions in thehands of engineers and designers. Yet, social science today can be practical like never before.

Academic institutions and researchers have only started to grapple with the implications of thisnew reality. By focusing on the role of the reputation system in creating thin trust, this paper offers a fresh perspective for a new and more applied role of social sciences. We will get back to this point in the conclusions.

[11] The distinction we make is relative. We do not claim that social science had no impact or that the tools at our disposal were free of consequences. Surveys, for instance, are extremely consequential (cite). Yet the migration ofour significant portion of social life to digital interface, created an explosion of data and an exponential growth in the ability to curate social realities with the strike of a keyboard.

We organize this chapter as follows. First, we will review the relevant literature on the topic of trust. The focus is on interpersonal trust rather than trust in institutions or generalized trust in others. We will then present evidence of how the reputation system is creating trust on these platforms. Most of the evidence comes from Parigi's previous work on platforms like Airbnb, Uber, and CouchSurfing. After presenting this evidence, we will explore the transformation thatthe concept of trust had undergone in terms of rationalization and examine the opportunities and risks that this new format of trust creates from social researchers and society. In conclusion, we will suggest a novel approach for scientific knowledge in the social sciences thataims at becoming applied.

TRUST IN NETWORKS

What is a trust? What does it mean to trust another person? Trust hovers between calculated action and blind faith. As Giddens notes, trust is necessary only in conditions of *incomplete* information, where the limits of existing knowledge and calculation, requires actors to suspendtheir disbelief and commit to a line of action that is inherently risky.[12] (Giddens 1990, 33). The partner to exchange, even one that has proved reliable in prior interaction, can always renege at the last minute. To complete such a transaction, the actors must, at some point, suspend their doubt and commit to the exchange. But the suspension of doubt is not equal to blind faith. In typical circumstances, the trusting actors rely on contextual cues (past experience, the context of the interaction, and the assessment of third parties' behavior) that turn trust into a reasonable if not fully calculated choice.[13]

[12] ANTHONY GIDDENS, THE CONSEQUENCES OF MODERNITY 33 (1990).
[13] GIL EYAL, THE CRISIS OF EXPERTISE (2019).

Given its centrality in social life, social scientists have studied and developed competing conceptions of interpersonal trust.[14] For rational choice theorists, people trust each other because of the benefits that trust generates.[15] Building on this approach while dispensing with the criticism that rationality requires perfect information,[16] some scholars have argued that trust emerges when the interests of the two parties engaged in the interaction are aligned.[17] On the contrary, other scholars have argued that trust is precisely needed when the parties' interests are unknown.[18] Finally, for students of culture, trust between people is the result of norms that shape society and get passed to individuals through institutions like the family and school.[19]

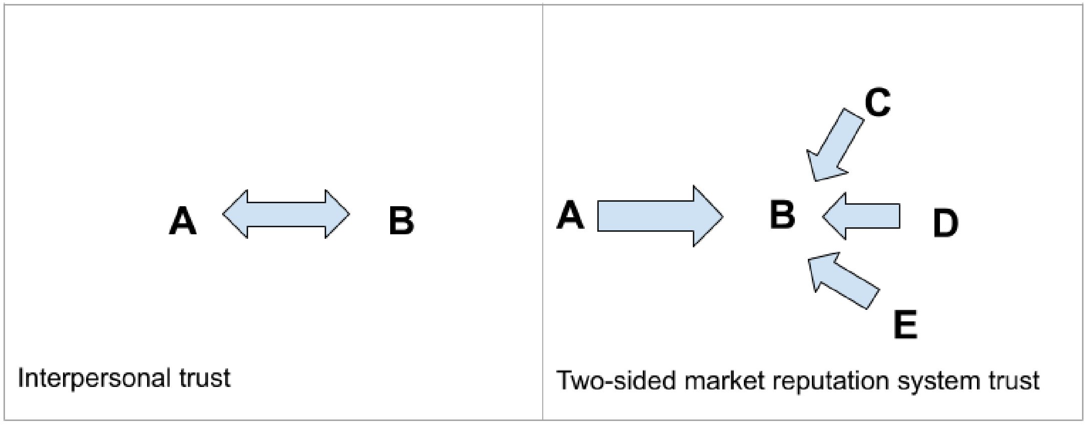
Mark Granovetter places the study of trust on empirical grounds by linking it to concretesocial networks: "You may trust that potential leader if there is a link or short chain of personal links to that person that conveys enough information to afford you some confidence that shewill act in a trustworthy manner."[20]

Granovetter's work clarifies that trust rests on the flow of information. Yet, he limits this flow to personal links and implicitly equates trust with in-person interactions. Personal links carry information and accountability and, because of that, can create a trust chain.[21] Trust networksin two-sided markets have a different structure than traditional trust networks (see Figure 1). Intraditional

[14] Karen S. Cook & Bogdan State, *Trust and Economic Organization*, *in* EMERGING TRENDS IN THE SOCIAL AND BEHAVIORAL SCIENCES 1 (2015).
[15] JAMES S. COLEMAN, FOUNDATIONS OF SOCIAL THEORY (1994).
[16] ROBERT GIBBONS, A PRIMER IN GAME THEORY (1992).
[17] RUSSELL HARDIN, TRUST AND TRUSTWORTHINESS (2002).
[18] PETER BLAU, EXCHANGE AND POWER IN SOCIAL LIFE (1986).
[19] FRANCIS FUKUYAMA, TRUST: THE SOCIAL VIRTUES AND THE CREATION OF PROSPERITY (1995).
[20] MARK GRANOVETTER, SOCIETY AND ECONOMY: FRAMEWORK AND PRINCIPLES (2017).
[21] Ferdinand Tönnies, Community and Civil Society (Margaret Hollis trans., 2001).

settings, trust results from repeated dyadic interaction between individuals (see left panel of Figure 1). Trust builds up incrementally, over time, and solidifies with the completion of each transaction. The repeated interpersonal nature of these transactions means that actors typically attribute trustworthiness to the actors they engage with. After a series of exchanges between A and B, A is likely to assume that B is trustworthy.

Figure 1: Configurations of trust



Interpersonal trust

Two-sided market reputation system trust

In two-sided markets actors are less likely to engage in repeated transactions. The informationthat actor A uses to determine the trustworthiness of actor B includes information posted by other actors (C, D, E, etc.) based on *their* previous interaction with B and depends on A’s valuation of the quality of the reputation system. Efficient information flow in the reputation system means that trust between strangers emerges almost instantaneously but results from the gradual accumulation of reviews of previous transactions.

Trustworthiness in this situation is impersonal in three senses. First, the information conveyed by previous reviews is not strictly related to B but to the interaction between B and C, B and D, B and E, etc. A must decide whether C’s review of B is credible and whether it reflects B or C or something in between. Second, the aggregation of past reviews through star ratings or otherwise means that when A determines that B is trustworthy or not, it does so by comparing B's ratings to other vendors on the platform. Trustworthiness thus becomes a relative property.In contrast with personal trust, which involves the presumption of reliability and involves the attribution of probity or honor,[22] trust in two-sided market emerges from comparing the ratingsof many others. Third, given that trustworthiness of an actor on the platform is tied with the credibility of the reputation system and the platform as a whole, trustworthiness is distributed between the two. Finally, given that two-side markets are domain-specific (Uber provides transportation, AirBnB provides hospitality), an actor's trustworthiness in a given market is noteasily transferable to other domains. This is why we call this new type of trust, "thin trust." Thintrust is impersonal and domain-specific.[23] It can effectively connect strangers and facilitate theirinteraction, but it is much narrower in meaning and scope than interpersonal trust.

The emergence of vast two-sided markets, in which trust relations connect a multiplicity of strangers, requires us to update Granovetter's perspective. We hold on to Granovetter's focus on information flows but note that reputation systems facilitate the flow of massive amounts ofpersonal data between agents and that this information flow sustains expansive trust networks decoupled from personal links. Knowledge of the most arcane things is now just a click away. In personal interactions, reputation systems have

[22] Giddens notes that this attribution renders trust psychologically consequential to the individual who trusts. *See* ANTHONY GIDDENS, THE CONSEQUENCES OF MODERNITY (1990).

[23] Unlike trust that is the result from the attribution of probity associated with a specific person, which is not strictly associated with particular domains, the attribution of trust in two-sided market platforms is associated with the particular service one offers. This domain-specific limits is, in part, a direct consequence of the fact that typical two-sided markets are specialized. Uber, for instance, provides transportation services, Airbnb specializes inhospitality, etc. and the reputations aggregated in one platform are not available in another.

distilled detailed and personal information in a digestible way designed for scaling and diffusion. Personal information used to require time to acquire. Now personal information about perfect strangers is available to participants of many online platforms as soon as they join the platform. Updating Granovetter's approach to incorporate technology means extending personal links to encompass online personal ties/information.

Online trust networks differ from Granovetter's face-to-face trust networks not only in scale and speed but also in their intermediation. Unlike spontaneously emerging face-to-face trust, trust in two-sided markets, Trust and Safety teams or divisions within many platforms, cultivateand curate the emerging networks. These teams' objective is to protect their users, but their main byproduct is trust. From this perspective, trust online is a network good that is actively being generated by users of a platform when they contribute with reviews and ratings and tech companies' workers when they make sure the feedback is authentic.[24]

For the most part, the main scope of Trust and Safety teams is to reduce fraud. Platforms attract many scammers who seek to exploit the vulnerabilities of a system that relies on mass participation is loosely supervised. Scams range from a host falsely advertising a property on Airbnb they do not own to elaborate fake accounts on Uber generating demands for reimbursement. Scammers are continually testing the network for vulnerabilities.

[24] While it may be the case that reputation systems online emerged by happenstance, their maintenance, controland evolution are essential part of two-sided markets. "When eBay launched, the biggest challenge was that consumers simply did not trust that they would getwhat they paid for. eBay quickly realized that without consumer trust, the system could not work. In response, eBay created the first Trust and Safety team, which was tasked with ensuring the trustworthiness of the eBay ecosystem." Amy J. Schmitz & Colin Rule, *Lessons Learned on eBay*, in THE NEW HANDSHAKE: ONLINE DISPUTE RESOLUTION AND THE FUTURE OF CONSUMER PROTECTION 33 (2017).

Containing fraud and malfeasance behavior is the primary goal of Trust and Safety teams in all two-sidedplatforms.

The containment and reduction of bad actors increase the reputation system's credibility and contribute indirectly towards increasing trust. The creation of badges and special statuses for users that passed carefully chosen milestones are also essential signals that facilitate trust between parties. An apt example is Airbnb's Super Host status on its platform. Airbnb reserves the status for hosts that meet specific criteria, and it is a signal of the host's trustworthiness. Interventions of this type directly create trust by extending the reputation of users. Thus, trust and safety teams indirectly create trust by containing bad actors and directly creating trust by expanding reputation signals.[25]

HOW ONLINE REPUTATION SYSTEMS GENERATE TRUST

The nature of online trust networks and their malleability creates opportunities for studies demonstrating what "thin trust" is all about. In a recent study, Parigi and his colleagues experimented to explore the extent to which the reputation system extended trust beyond homophily.[26] Homophily, one of the few constant behaviors of social life,[27] is the tendency to interact and trust others who are similar.[28] The researchers set up an online experiment basedon the widely used investment game that simulates actors' behavior in two-sided markets. In two-sided markets, participants decide whom to trust based on the information displayed about the unknown alter.

[25] Notice that the expansion of signals is limited to a specific platform. The Super Host status does not apply to adriver for Lyft, for instance.

[26] Abrahao et al., *supra* note 4.

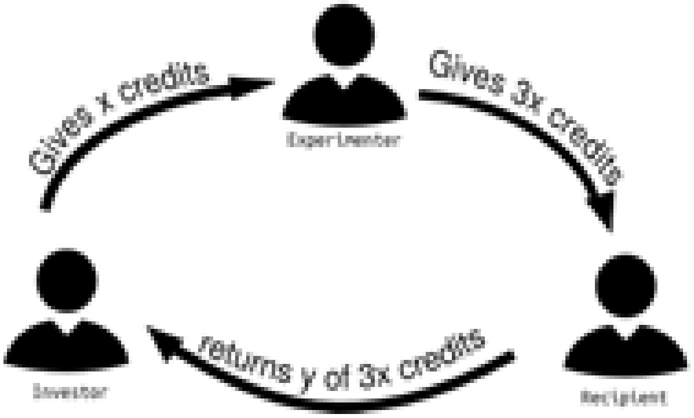
[27] PETER BLAU, INEQUALITY AND HETEROGENEITY (1977).

[28] Miller McPherson, *An Ecology of Affiliation*, 48 AM. SOCIOLOGICAL REV. 519 (1983).

Similarly, in the investment game, users have to decide whom to place trust based on limited information.

The investment game is a single-shot game where participants decide how many credits to invest in a recipient. Recipients receive three times that amount and may cooperate or defect when determining how many credits to return to the investor. The figure below shows astylized version of the game:[29]

Figure 2: The standard trust game.



For example, if a participant decided to invest 5 points, the recipient will receive 15 points and choose how many points to return. Parigi and his colleagues led all participants in the experiment to believe that they were randomly assigned the role of investor and instructedthem to play with five other Airbnb users cast to recipients' role. In reality, the recipients weresynthetic profiles that the researchers concocted. As investors, participants received 100 points and had to decide how to allocate them. The experiment involved almost 9,000 Airbnb users in the United States.[30]

[29] Will Qiu et al., *"More Stars or More Reviews?"*, *in* PROCEEDINGS OF THE 2018 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (2018).
[30] Abrahao et al., *supra* note 4.

The synthetic profiles had different demographic characteristics (gender, age, marital status, and U.S. region) and different reputation levels, all varied in a structured way. For the demographic characteristics, profiles were located at various social distances ranging from matching all the participant attributes to differing in all the attributes. A profile at (social) distance0 had demographic characteristics that fit the participant's profile, while a profile at a distance of 4 was the most dissimilar. To illustrate, imagine a male player from California, not married, and 40 years old. The profile at distant 0 will have all the same characteristics of the player, while the profile at a distance of 4 will be all different, i.e., a female from New York, married in her 60s.

The 5th profile was identical to the profile at a distance of 4 but had a different reputation from all others. The experiment had two conditions—one in which the 5th profile had a worse reputation than the previous four (world 1, Figure 2 left panel) and one in which she had a better reputation (world 2, Figure 2 right panel).

Figure 3: Main results of an online experiment

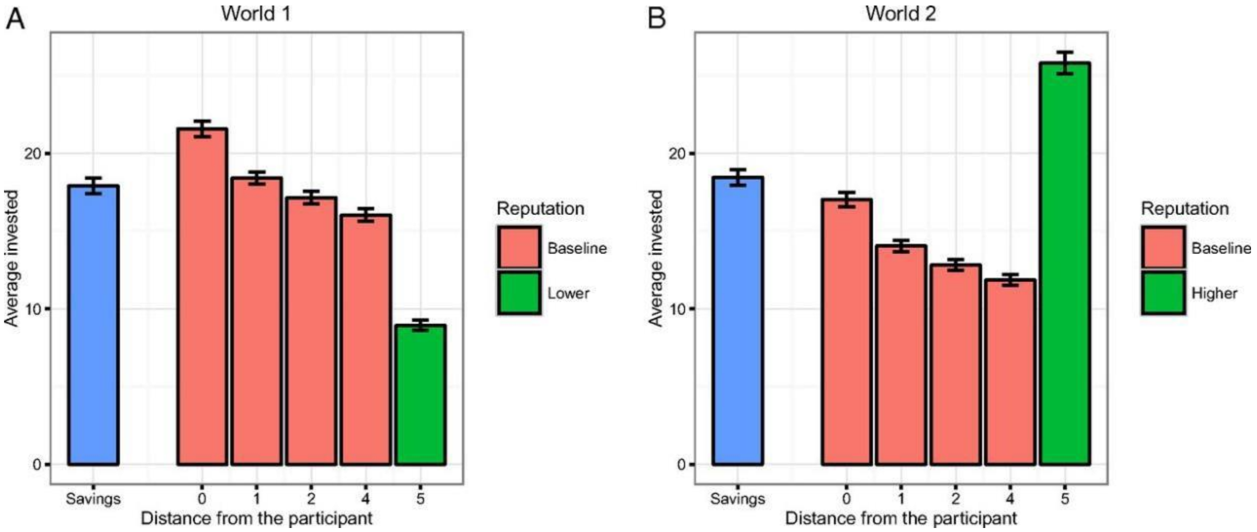


Figure 3 presents the main result of the research. The x-axis plots the social distances on the profiles, 0 to 5. As previously explained, distance 0 means a profile that is identical to the participant; distance 1 means a profile with one characteristic different from the participant, and so on. Distance 5 (green bar) represents a different profile from the participant and has thesame characteristics as distance four but a different reputation. Note that there is no distance 3because of the difficulty in interpreting 3-way interaction effects. The y-axis plots the average amount of points invested. In both panels, the blue bar shows the average amount of points not invested (, i.e., saved).

On the left panel, the effect of homophily is almost self-explanatory. The more socially distant a profile was from the participant, the lower the number of points invested. In other words, the left panel confirms that people trust others who are similar to themselves. Homophily works online as it works offline. Note in this condition, world 1, that the 5th profile paid an extra penalty caused by his worse reputation—the decrease in points invested in him compared to the profile at a distance 4 is large and significant.

Focusing on the right panel, or the condition in which the most diverse profile (green bar) has a better reputation than all the other profiles. The plot shows a dramatic increase in trust. A positive reputation significantly extends trust beyond the effects of homophily (still visible in the declining trust in the red bars as social distance increases). After controlling for various factors and considering the complex experiment's dependencies, the researchers concluded that the reputation system significantly extended trust towards different others.[31] The reputation system makes possible the circulation of trust in two-sided markets.

[31] *Id.*

Participants/investors interpret the ratings and reviews as signals for trustworthiness, and because of that, engage in the exchanges the platform offers.

Looking at the Stars

In most platforms, reputation systems have two components, ratings, and reviews. Ratings areusually expressed on a 5-star scale, while reviews consist of comments that users left about their experiences. Using the same data described above, Qiu et al.[32] separated the impact of the two parts of the reputation system on perceptions of trustworthiness.

In the experiment above, participants were exposed to a star condition—4 or 5-star ratings—and a review condition—a low number of reviews (1-3) or a high number of reviews (11-50). While the difference between 4 stars to 5 stars may appear limited, at first sight, it mirrors a reality in which the overwhelming majority of ratings available on two-sided marketsare positive.

Qiu et al. compared the reputation system's impact by focusing on profiles at distance 4 and distance 5. Both profiles have the same demographic characteristics but a different reputation.In particular, they fit the following mode:

$$[1]\quad Y_{ij} = \mu + \alpha_j + \beta_1 s_i + \beta_2 r_i + \beta_3 (s_i \times ri) + \beta_4 w_i + e_{ij}$$

Where Y_{ij} is the predicted investment amount for profile *i* by subject *j*. μ is the global interceptat a star rating of 4 and Low Review count in world 1, and α_j are random intercepts to account for individual variations. β_1 is the profile level estimate of having 5 stars, β_2 is the profile level estimate of having High review counts,

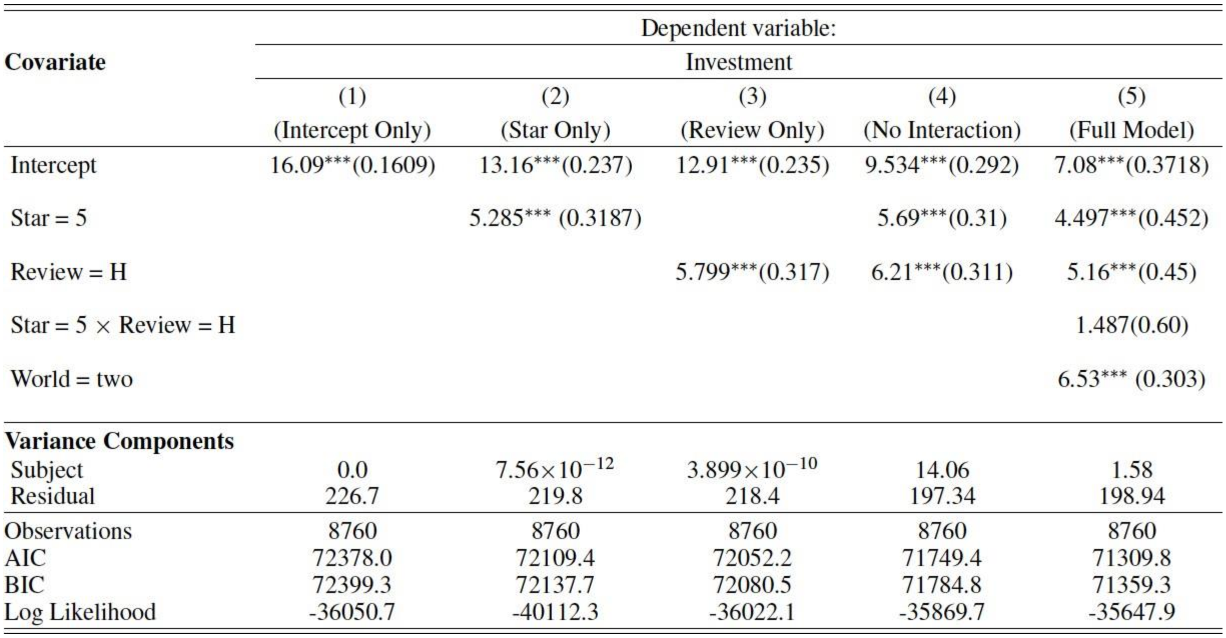
[32] Qiu et al., *supra* note 29.

and β_3 is their estimated interaction effect. β_4 is the estimate of a profile *i* being placed in world two, and $e_{i\,j}$ is the random error.

Star rating (s) is a factor variable with two levels (4 stars or 5 stars), and review condition (r) also has two levels (Low or High). Because there were multiple measurements of investments per participant (each participant invested p4 and p5), the measured investments are correlated. To account for this, the model nested profile investments within subjects by fitting simple random subject-level intercept α_j. The model also does not include an explicit term for social distance because we confined our analysis to observed investments between p4 and p5 only, who share the same distance to the subject but different reputations.

Table 1 summarizes the results from Qiu et al. The table shows estimated fixed effect coefficients for five models: (1) an intercept only model, (2) star rating only model, (3) reviewcount only model, (4) additive model of star and review, and lastly a full model with (5) both additive as well as interaction terms.

Table 1: Multi-level Model Estimates of Star Ratings and Review Counts Components on Investment



Covariate	Dependent variable:				
	Investment				
	(1)	(2)	(3)	(4)	(5)
	(Intercept Only)	(Star Only)	(Review Only)	(No Interaction)	(Full Model)
Intercept	16.09***(0.1609)	13.16***(0.237)	12.91***(0.235)	9.534***(0.292)	7.08***(0.3718)
Star = 5		5.285*** (0.3187)		5.69***(0.31)	4.497***(0.452)
Review = H			5.799***(0.317)	6.21***(0.311)	5.16***(0.45)
Star = 5 × Review = H					1.487(0.60)
World = two					6.53*** (0.303)
Variance Components					
Subject	0.0	7.56×10^{-12}	3.899×10^{-10}	14.06	1.58
Residual	226.7	219.8	218.4	197.34	198.94
Observations	8760	8760	8760	8760	8760
AIC	72378.0	72109.4	72052.2	71749.4	71309.8
BIC	72399.3	72137.7	72080.5	71784.8	71359.3
Log Likelihood	-36050.7	-40112.3	-36022.1	-35869.7	-35647.9

The results show a significant increase in average investment received when a profile goes fromhaving 4 stars to 5 stars (~ 4.5 more credits)) as well as going from low review to high review (~ 5.16 more credits). Their interaction (β_3 = 1.487) is not statistically significant. Having a 5- star rating and lots of reviews does not significantly increase trust in the profile. Either one of the two conditions suffice. Finally, the estimates are stable even when we consider world differences. The researchers summarized their findings: "for a profile, the effect of going from having 4 stars to 5 stars on the number of credits is equivalent to the effect of going from having only 1-3 reviews to having at least 11 reviews on average."[33]

The arbitrariness of using a 5-star scale for ratings and the peculiarities of many of the comments left on these platforms has made many observers think that the reputation system isan ancillary add-on to many websites. This analysis shows that the reputation system is crucial for creating extended trust. These platforms' users interpret both the ratings and the reviews as signals for trustworthiness. These signals represent thin trust because the judgment of trustworthiness is both numerical and relative. Users do not engage with the profile characteristics but with the profile in relation to other profiles.

MODELING AND THINNING TRUST

The capacity to measure trust has been the holy grail of trust scholars for many years because trust is essential for economic growth,[34] the health of institutions,[35] and individual well-being.[36]

[33] Qiu et al., *supra* note 29, at 7.

[34] Stephen Knack & Philip Keefer, *Does Social Capital Have an Economic Payoff? A Cross-Country Investigation*, 112 THE Q.J. ECON. 1251 (1997).

[35] John F. Helliwell & Robert D. Putnam, *Economic Growth and Social Capital in Italy*, 21 EASTERN ECON. J. 295 (1995).

[36] ERIC M. USLANER, THE MORAL FOUNDATIONS OF TRUST (2002).

Yet, measuring trust has proven elusive due to the concept's subjective and fleeting nature. More importantly, even when researchers successfully measure trust, the techniques for doingso—detailed attitudinal surveys—are imprecise, costly, and therefore quite rare.

The penetration of technology in many aspects of life changed this and made it possible to accumulate data on private interactions that were previously unthinkable. The optimization of the virtual spaces where these interactions occur allows measuring trust using behavioral data rather than relying mostly on costly attitudinal surveys. To the extent that technology has entered many more contexts of contemporary life, from walking your pet, to hosting people, to suggesting potential romantic partners, it has created a world that is amenable to digital experimentation, measurement, and optimization.[37] While trust in your loved one may not be measurable, the trust that circulates on a platform like Airbnb is. A progressively more digital world is also a more quantifiable world. It is also a world where trust can be carefully designed.

Barbosa et al.'s work illustrates these new opportunities.[38] The researchers developed a data triangulation process by which they collected data first using an online experiment very similarto what we described above. The experiment provided them with a measure of the trusting behavior of about 5,000 Airbnb users. Using machine learning, they then create a model to identify low, middle, and high trust levels. The model identified actions taken on the platform, i.e., logged behavior, that correlated with trust levels.

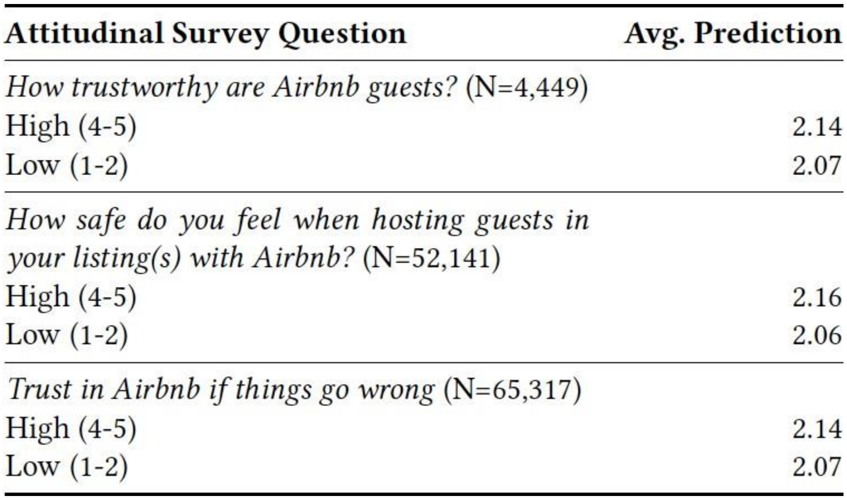
[37] Xiao Ma et al., *When Do People Trust Their Social Groups?, in* PROCEEDINGS OF THE 2019 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (2019).

[38] Natã M. Barbosa et al., *Designing for Trust: ABehavioral Framework for Sharing Economy Platforms, in* PROCEEDINGS OF THE WEB CONFERENCE 2133 (2020).

Among the actions that correlate positively with trust among guests were: (a) reading the full description of the host profile and (b) the length of the communication with the host. Factors that were negatively associated were: (a) reading reviews in full and (b) level of engagement with other guests generated content.[39] Among hosts, the positively correlated actions were: (a) total engagement (i.e., time spent in the app), (b) prior requests of engagement, while the negatively correlated behaviors were: (a) the number of communication exchanges and (b) number of rejections.[40]

In a second step, Barbosa et al. validated their trust model by changing the target variable. Thatis, they used the model to predict answers to a set of trust questions provided by Airbnb hosts and guests. This sample was much larger, about 200,000 respondents, and completely independent of the sample that participated in the experiment. Most relevant, the dependent variables for the predictions were survey questions rather than behavioral variables. The table below summarizes the results of this second step for Airbnb hosts:

Table 2: Triangulation of attitudes and experimental data.



Attitudinal Survey Question	**Avg. Prediction**
How trustworthy are Airbnb guests? (N=4,449)	
High (4-5)	2.14
Low (1-2)	2.07
How safe do you feel when hosting guests in your listing(s) with Airbnb? (N=52,141)	
High (4-5)	2.16
Low (1-2)	2.06
Trust in Airbnb if things go wrong (N=65,317)	
High (4-5)	2.14
Low (1-2)	2.07

[39] *Id.* at 2138.
[40] *Id.*

For all the questions, the high trust group model's average predictions are higher than for the low trust group, and the differences are all statistically significant. In other words, the model did a good job in predicting trust beliefs of users from a different sample. This step connected the behavioral model to the attitudinal model. Guests that exhibited a particular set of behaviorcorrelated with, say, low levels of trust also had low trust beliefs towards hosts. Finally, Barbosa et al. trained a neural network to predict low, medium, and high trust levels for hosts and guests. The applied the neural network model to classify another batch of hosts and guests.

Their model performs very well in predicting users with low levels of trust, while it is lessaccurate for users with high trust levels.[41]

The procedure we describe above illustrates the technological capacity to measure and intervene in the production of trust and its subsequent thinning. The measurement of trust, which used to be costly and rare, is now accomplished at a relatively low cost and applied to the entire population of users.[42] Once done, the researchers can intervene and study how various changes to the platform change trusting behavior, again, at a scale previously unimaginable. Note that the trust we enact and work on in this setting is thin. Barbosa and his colleagues do not attempt to model and work upon a thick interpersonal relation but merely toalter user behavior on a single platform.

The measurement and thinning of trust rely on the reputation system's technology. In the absence of such technology, trust

[41] Barbosa et al., *supra* note 35.

[42] In the past, measurement of trust relied primarily on attitudinal surveys. Such surveys were able to measuregeneral tendency to trust others, for instance. But attitudinal surveys are costly, rare, and lack the specificity required to intervene in practical action.

remains mostly inoperative. One of the early two-sided marketplace platforms, CouchSurfing,[43] is a case in point because it never implemented a reputation system. Instead, it relied heavily on its users posting detailed descriptions of themselves and their interactions. In describing the world of CouchSurfing, Patrica Marx—a writer for the New Yorker—wrote:

> Upon joining CouchSurfing, you are instructed to compose an online profile, delineating your philosophy and mission, the skills you can teach others, your favorite music, movies, and books, and so much else that you might as well be applying to college.
>
> Members also post photographs of themselves, sometimes hundreds of them.[44] Without the reputation system's technology, trust on the CouchSurfing platform never became measurable and thin, in the sense described above. It remained a deeply personal experience rooted in knowing the other. This is what Paula Bialski wrote about her first hosting experience on CouchSurfing: "He [the guest] would speak, and I would often listen. It was thefirst time I ever invited a stranger into my home, and the first time I ended up speaking to a stranger until the late hours of the night.[45]

It may be useful to think about the transformation of trust as an instance of rationalization. According to Max Weber, rationalization is the process through which more and more spheres of life become subject to calculation, measurement, and control

[43] Established in 2003, it reached 1 million users in 2009 and would go on to sign up more than 10 million users by2015. *See* Coca Nithin, *The Improbable Rise and Fall of Couchsurfing*, THE TRAVEL CLUB (June 12, 2015), https://www.thetravelclub.org/articles/traveloscope/698-the-improbable-rise-and-fall-of- couchsurfing/. Notwithstanding, the platform was never profitable and became mired in several legal controversies. The platform is now a marginal player in the two-sided marketplace segment. Yet CouchSurfing remains important for its pioneer role in creating a different way to travel.

[44] Patricia Marx, *You're Welcome*, THE NEW YORKER (Apr. 16, 2012), https://www.newyorker.com/magazine/2012/04/16/youre-welcome.

[45] PAULA BIALSKI, BECOMING INTIMATELY MOBILE (2012).

(Weber, Science as a Vocation).[46] While rationalization brings a leap in efficiency, Weber was deeply ambivalent about the process. Alongside increased efficiency, Weber noted that rationalization ushers a process of disenchantment. Disenchantment refers to the semiotic changes that result from the application of measurement and calculation to actions and situations that were previously less rationalized. Measurement, in other words, does more than reflecting the state of the world. It also changes its meaning. For instance, the measurement of economic activity, like gross domestic product (GDP), had not only rendered a previously abstract entity (the economy) visible and actionable in ways previously unimaginable, it also fundamentally shifted how policymakers think about "the economy." Once developed and implemented, economic growth became an end of its own, and policymakers now design interventions designed to boost GDP growth.[47]

Similarly, the development and widespread implementation of intelligence tests had altered the meaning of wisdom. From a holistic attribute of a person, appreciable in conversation or through an in-depth acquaintance, being smart is gradually reduced to solving a series of relatively meaningless multiple answer questions at a given speed.[48] Note that in both examples, the development of measurement procedures does little to clarify the terms' ambiguities. The economy remains an abstract concept whose boundaries are imprecise, andwisdom remains an elusive and confusing attribute. More pointedly, the development of measurement procedures to capture "the economy" or "intelligence" reduced their meaning. From wholesome concepts that resist measurement but convey deep

[46] Max Weber, *Science as a Vocation*, 87 DAEDALUS 111 (1958).
[47] Timothy Mitchell, *Economentality: How the Future Entered Government*, 40 CRITICAL INQUIRY 479 (2014).
[48] NIKOLAS S. ROSE, GOVERNING THE SOUL: THE SHAPING OF THE PRIVATE SELF (1999).

meaning, these concepts' measurement turned them measurable but almost meaningless.[49]

The measurement and modeling of trust, especially integrating these models into two-sided markets, is affecting a similar transformation. Trust remains an abstract concept, slippery and full of ambiguities. Still, now platform operators can respond in real-time to challenges and mistrust and actively optimize two-sided markets to generate trust. The type of trust that emerges through such intervention is different from the interpersonal trust that actors skillfully develop through repeated interpersonal exchanges. In place of interpersonal reciprocal trust, this new type of trust results from the accrual of reviews of past transactions with third parties. This trust, as we have seen, is mostly impersonal. It is not based on past interaction with the trusted actor but on others' experiences. It is based on the position of the trusted actor relativeto other actors on the platform.

Importantly, this trust is also thin because it is domain-specific. This specificity of trust is, in part, merely a function of the organization of reputation systems, which are nested within specific two-sided markets (Uber, AirBnB, Amazon, etc.). But it is also a byproduct of the fact that this type of trust is impersonal and detached from the actors' actual past experiences.

Domain-specific trust is not a new phenomenon. We typically trust our doctor's advice on matters that pertain to health but will be quite cautious when it comes to assessing her stock purchase advice. We trust our lawyer (always a mistake) for legal advice, not health matters. Yet, the domain-specific trust that

[49] Researchers sometimes argue that intelligence is precisely what we measure in psychometric test. *See* Claude S. Fischer et al., *Understanding 'Intelligence'*, in INEQUALITY BY DESIGN: CRACKING THE BELL CURVE MYTH 22 (1996).

reputation systems generate does not depend on diplomasor other forms of credentialed expertise. Relatively thin trust relies on the labor of previous reviewers.

Finally, whereas interpersonal trust relies on the parties' skilled interaction, the mediating role that reputation systems play in creating and sustaining this type of trust, means that some of the skill involved in creating the trust does not reside between the parties that exchange goods and services. Instead, this skill is appropriated, in part, by the Trust and Safety teams or otherplatform operators that continuously experiment and optimize their systems. To the extent that this is the case, thin trust operates "under the hood" or outside the consciousness of the involved actors. The development of reputation systems powers a leap in the scale of trust relations, but it leaves the parties to the exchange without a clearer understanding of the conditions within which they live and act.

TOWARD APPLIED SOCIAL SCIENCE

This chapter explores the development of two-sided markets, focusing on how technologically sophisticated reputation systems foster the creation of thin trust between actors on those platforms. The chapter also calls attention to a new frontier for the social sciences. In the past, social research was an academic pursuit. To the extent that social scientists found their way to industry, their roles were typically marginal and confined to consumer behavior studies throughsurveys or focus groups. However, the digitization of everyday life creates entirely new possibilities for the integration of social science and business. Along with these possibilities come new ethical questions and risks. This last section returns to the issue of trust to explore these new possibilities and dangers.

Data exists on things that used to be beyond the reach of quantification and experimentation. The range of questions that social scientists can now ask has expanded. More importantly now it exists as a mechanism through which interactions can be planned and their consequences measured. Such is the nature of socio-technical systems, and social scientists could be a part of the solutions that get designed.

Social phenomena depend on the interaction of multiple actors. Until recently, it was practicallyimpossible to intervene and experiment within such interactions.[50] The digitization of social life changes allows social scientists to experiment on a very large scale. Importantly, the realism of these interventions, since we operate on the same platforms and interfaces actors use in their everyday life and in the same settings, is very high.

The digitization of social life presents social scientists with an exciting research frontier. More than that, the mediated nature of online platforms effectively allows social scientists not only to study but to intervene and curate social interactions at scale. A digitized social space means aspace where operators can plan and measure every interaction. The analogy with urban planning is apt. In this newly digitized space, operators can experiment and optimize interactions in the same way that urban planners design urban spaces and traffic flow, but with far better efficiency.

If trust could be measured and modeled, it can also be manipulated in more invasive ways. For instance, in 2014, Uber launched a carpooling service on its app, allowing users to share a ride. However, putting strangers in the same car is a tricky socio-technical feat. Part of the challenge was to match riders to correct routes efficiently. But Uber quickly discovered that bad matchingof

[50] Social psychologists attempted to do that by treating the individual as the locus of the experiment, but the *raison d'être* of the social sciences is to study relations between individuals.

riders could result in unpleasant altercations and unexpected challenges to drivers.[51]

Friction in the interface between service providers is nothing new, but in the past, companies had limited ability to respond to these breaches of trust in real-time. Uber or its likes could have intensified background checks and ban problematic drivers or passengers, and it could introduce new rules of behavior in rides. But powered with a good trust model, platform operators could have introduced a whole new roster of interventions to prevent the problem.With better modeling of drivers and consumers, Uber now could have prevented matchings ofincompatible riders or drivers, it could have identified difficult times or areas of service, and itcould have created changes in the app itself to help consumers report challenging encounters in real-time (which could be used to further optimize the model). None of these possibilities existed before, and for sure, Uber did not deploy the solutions we cursory mentioned above . Yet, they remained possible and platform operators could test whether any of these solutionsworked, and to do that in an extremely short interval. A lack of trained social scientists is the main reason why these solutions were not tested.

Uber and similar platforms operating in two-sided markets are modeling interactions. The consequences of their products do not remain confined within the virtual worlds they create. Instead, their products intervene and alter people's social interactions. Two-sided markets have created opportunities for social scientists to measure and design social interactions at a scale not previously possible. While exciting, these developments pose challenging ethical questions. Using trust as an example, users identified as a

[51] Kiana Cornish, *'Ride from Hell': Carpooling in the Age of Uber Can Be…Awkward*, WALL ST. J (Dec. 6, 2018), https://www.wsj.com/articles/ride-from-hell-carpooling-in-the-age-of-uber-can-beawkward-1544112559.

low trust based on their actions on the platformcould be exposed to different conditions to win back some of their trust. Yet, the model upon which trust levels are predicted ignores the personal reasons why users may have different levels of trust. Intervening to bypass barriers to trust may become a manipulation that reduces individual choices. Informing users about the potential existence of such models is only a first step in protecting users' freedom.

A better and more systematic approach to address ethical questions would require the platforms to leverage social scientists' expertise in designing and planning products like a reputation system. Social scientists are uniquely capable of understanding the impact of socio- technological systems. For example, when Nextdoor was trying to find a solution to racist comments on its platform, they hired Jennifer Eberhardt, a Stanford social psychologist. Nextdoor CEO described Eberhardt's work:

> The basis of her research is around something she calls decision points. If you make people stop and think before they act, they probably won't do the racist things that theydo." Today, if you post in the crime section and decide to use race to describe a person, the platform makes you fill in two other characteristics. This simple intervention reduced racist posts by 25% in 2016.[52]

However, social scientists were not included in designing the app from its beginning; neither were they part of its measuring and monitoring. Instead, Nextdoor stumbled upon the solution after other approaches failed and the community faced significant strife. The penetration of technology into more life domains has created the space for applied social science.

[52] Pendarvis Harshaw, *Nextdoor, the social network for neighbors, is becoming a home for racial profiling*, SPLINTER (Mar. 24, 2015, 10:02 AM), https://splinternews.com/nextdoor-the-social-network-for-neighbors-is-becoming-1793846596.

INTERNET GOVERNANCE AND HUMAN RIGHTS IN A MINOR KEY: AN ANTHROPOLOGICAL PERSPECTIVE

Baron Pineda[*]

This paper explores the intersection of human rights and internet governance with the field of anthropology. Regimes of internet governance and platform content moderation are carried out on a global scale. They engage with cross-cultural issues that are central to anthropology, such as cultural relativism and legal pluralism. The discipline of anthropology has a long tradition of skepticism towards the international human rights movement. However, in recent decades many anthropologists have developed approaches to universal human rights that have overcome their natural objections and concerns. An examination of the ways that rights-focused anthropologists have addressed these concerns provides a productive way to refine and fortify human rights approaches to internet governance. This paper illuminates points of convergence between a rights-focused anthropology and the specific approaches to internet governance that have been developed in circles outside of anthropology.

INTRODUCTION

In recent decades, internet governance has emerged as a new area of study. With the rise of internet-based communication, this field is dedicated to understanding the challenges posed by new ways that people and institutions interact within the World Wide

[*] Professor, Department of Anthropology, Oberlin College.

Web.[1] Classic themes in law and public policy (such as privacy, defamation, antitrust, security, public safety, surveillance, corporate responsibility, and copyright infringement) are well represented in this area of study, and the stakes are high. However, as we think about the rise of the internet and the corresponding challenges of internet governance, it is important to take note of the ways in which the legal and policy framings that present themselves in this area resonate deeply with heavily studied subjects in anthropology and other humanistic social sciences.[2] These include subjects such as cultural relativism/universalism, critiques of Eurocentrism, new forms of colonialism and imperialism, technologies of control and power, the nature of freedom,[3] as well as utopian and alternative forms of democratic participation and citizenship.

Consider the following list of new hybrid (i.e. machine-human) words and phrases that are closely related to internet governance and have entered the modern lexicon, both in popular usage and as analytical terms. These terms echo a previous generation of anthropological debates and redirect contemporary ones: “data colonialism,” “internet freedom,” “cybersovereignty,” “algorithmic racism,” “computational propaganda,” “artificial intelligence,” “netiquette,” “data nationalism,” “surveillance capitalism,” “big data,” “open access,” “digital divide,” “net neutrality,” “social computing,” “machine learning,” “cyberbullying,” and “online harassment.” Terms like these pair the

[1] Milton L Mueller & Farzaneh Badiei. *Inventing Internet Governance: The Historical Trajectory of the Phenomenon and the Field, in* RESEARCHING INTERNET GOVERNANCE: METHODS, FRAMEWORKS, FUTURES 63 (2020).

[2] Anthropologist Anna Cristina Pertierra describes the “four basic premises of Anthropology” as the following: 1) cultural relativism 2) holism 3) “deliberate esoterism”—that is, attention to the marginal—4) ethnographic. ANNA CRISTINA PERTIERRA, MEDIA ANTHROPOLOGY FOR THE DIGITAL AGE 5-8 (2018).

[3] LAURA DENARDIS, THE INTERNET IN EVERYTHING: FREEDOM AND SECURITY IN A WORLD WITH NO OFF SWITCH 187 (2020).

worlds of computer networks with classic social science preoccupations relating to social life.

How and why should we study the discourses and debates that have emerged around these themes from an international ethnographic point of view? For example, is social media content moderation an exercise in "moral imperialism" given the generation of these policies in the metropoles and application in the global peripheries?[4] Can and should content moderation standards be adapted to the local cultural and social contexts in which they are applied? What about the Utopian language of the early internet that envisioned the "World Wide Web"[5] as a "place" that would offer people the potential for freedom from the constraints of national and corporate power? How and why should we study the emergent discourses of globalism and deterritorialization that the rise of the internet has precipitated? These are questions that lie at the intersection of anthropology, human rights, and internet governance.

One way to address these questions from an anthropological perspective is to examine the engagements that anthropology has made with the human rights movement.[6] There are many reasons that considering the postwar human rights movement alongside the emergence of internet governance in the digital age is a productive undertaking. Both represent attempts at creating new forms of global

[4] MORAL IMPERIALISM: A CRITICAL ANTHOLOGY (2002).

[5] TIM BERNERS-LEE, WEAVING THE WEB: THE ORIGINAL DESIGN AND ULTIMATE DESTINY OF THE WORLD WIDE WEB (2009).

[6] I have in mind an expansive definition of the "human rights movement" that combines: 1) the notion of rights that extend to all humans regardless of citizenship 2) the advocacy networks that ground their work in this notion and 3) the formal institutions of international human rights law that anchored by the Universal Declaration of Human Rights (UDHR) and subsequent UN Human Rights conventions. Distinguishing the broader term, "human rights movement," from the more specific term, "human rights law," is common in the scholarship of human rights. *See* ANDREW CLAPHAM, HUMAN RIGHTS: A VERY SHORT INTRODUCTION (2015).

governance.[7] Both appeal to rhetoric of the diminished salience of national borders and the harms of unqualified national (or corporate) sovereignty.[8] Both employ "constitutionalist" approaches—lists of rights meant to constrain the abuse of power.[9] Both attempt to intervene in geopolitical conflicts that are talked about on a "civilizational" scale—e.g., consider the tense East-West dynamics of the "The Great Firewall of China."[10] Both are initiatives that struggle with the reality of "American exceptionalism" and its "unique mission to transform the world."[11]

Anthropology as a discipline has a long tradition of skepticism regarding the international human rights movement for many reasons including the contention that the movement (in many of its variations) fails to live up to its universalist pretensions given the Western dominance and eurocentrism of its foundation and institutionalization. However, in recent decades many anthropologists have developed approaches to universal human rights that have overcome their natural objections and concerns[12]—going "from skepticism to embrace" in the words of legal scholar Karen Engle.[13] My contention in this article is that an examination

[7] Monika Zalnieriute & Stefania Milan, *Internet Architecture and Human Rights: Beyond the Human Rights Gap*, 11 POL. & INTERNET 6 (2019).

[8] MILTON MUELLER, WILL THE INTERNET FRAGMENT?: SOVEREIGNTY, GLOBALIZATION AND CYBERSPACe (2017).

[9] Lex Gill et al., *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*, BERKMAN CENTER RSCH. PUBL'N NO. 2015-15 (2015).

[10] RONALD DEIBERT, RESET: RECLAIMING THE INTERNET FOR CIVIL SOCIETY (2020); PEN America, *Forbidden Fees: Government Controls on Social Media* (2018), https://pen.org/wp-content/uploads/2018/06/PEN-America_Forbidden-Feeds-report-6.6.18.pdf.

[11] AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS (2005).

[12] MARK GOODALE, SURRENDERING TO UTOPIA: AN ANTHROPOLOGY OF HUMAN RIGHTS (2009).

[13] Karen Engle, *From Skepticism to Embrace: Human Rights and the American Anthropological Association from 1947-1999,* 23 HUM. RTS. Q. 536 (2001).

of the ways that rights-focused anthropologists[14] have addressed these concerns provides a productive way to refine and fortify human rights approaches to internet governance. In the process I will illuminate points of convergence between a rights-focused anthropology and specific approaches to internet governance (and social media content moderation) that have been developed in circles outside of anthropology.

CONTENT MODERATION AND SOCIAL MEDIA: RIGHTS AND "MERE WANTS"

Facebook, WeChat (China), Vkontakte (Russia), Twitter, and the rest of the social media platforms are engaged in an international law-like exercise when they establish "content moderation" rules for how user-generated content will be regulated on a global basis. They are attempting to establish a single set of standards that will be used to screen content in order to "facilitate cooperation and prevent abuse."[15] Internet reformers have appealed to human rights law as a set of mechanisms with which to address the problems associated with the centrality of social media in contemporary life. Proponents of a "rights-oriented regulation" promote the fortification of legal and political remedies that are built around those articles of the United Nations International Covenant on Civil and Political Rights that pertain to freedom of speech,

[14] For the purposes of this essay, I define "rights-focused anthropologists" as anthropologists who orient their work around struggles with the promise and limitations of human rights rhetoric and institutions. For an influential elaboration of this tradition, see Ellen Messer, *Anthropology, Human Rights and Social Transformation, in* TRANSFORMING SOCIETIES, TRANSFORMING ANTHROPOLOGY (1996).

[15] James Grimmelmann defines moderation in online communities as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 47 (2015).

particularly Article 19.[16] In the words of David Kaye, legal scholar and former UN Special Rapporteur on the promotion of freedom and protection of the right to freedom of opinion and expression, "It's time to put individual and democratic rights at the center of corporate content moderation and government regulation of the companies."[17]

Content moderation is the area of the broader world of internet governance that most obviously collides with the classic anthropological concerns with culture and cultural relativism. Seemingly more culturally sterile issues such as global network security and espionage fall under the umbrella of internet governance, but because content moderation consists of evaluating the details of user-generated content, it inevitably begs the question of who is doing the judging and on what basis. The basis on which a given norm is applied is a classic concern in anthropology that is often referred to as cultural relativism. The Oxford English Dictionary gives the following definition of the term: "The theory that there are no objective standards by which to evaluate a culture and that a culture can only be understood in terms of its own values and customs."[18]

Typically, internet companies have two ways of setting the ground rules for what will be acceptable on platforms—terms of service and community guidelines.[19] Terms of service are set up as contracts that establish rules and obligations between platforms and

[16] DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET (2019); Michael Karanicolas, *Squaring the Circle Between Freedom of Expression and Platform Law*, PITTSBURGH J. TECH. L. & POL. 177 (2020).

[17] KAYE, *supra* note 16. at 17.

[18] *Cultural Relativism*, OXFORD ENGLISH DICTIONARY, https://www.oed.com/view/Entry/45742?redirectedFrom=cultural+relativism#eid129084834 (last visited Jan. 2, 2021).

[19] JAMILA VENTURINI ET AL., TERMS OF SERVICE AND HUMAN RIGHTS: AN ANALYSIS OF ONLINE PLATFORM CONTRACTS (2016).

their users. Community guidelines are didactic and aspirational documents. They lay out "the platform's expectations of what is appropriate and what is not," and announce "the platform's principles, and list prohibitions, with varying degrees of explanation and justification."[20] All of the major global platforms commit themselves to monitoring and promoting the community guidelines which they publish. Tarleton Gillespie in his study of content moderation notes that these are "strikingly similar."[21] How these documents construct on a global scale what will be considered normal vs. abnormal, polite vs. offensive, respectful vs. sacrilegious, or tolerant vs. racist is a difficult exercise. Critics contend that the major platforms have failed to, in the words of Facebook's Mark Zuckerberg, "develop the social infrastructure to give people the power to build a global community that works for all of us."[22]

Facebook has an elaborate set of rules called "Community Standards" that they use to regulate speech on the platform.[23] These are divided into six categories: Violence and Criminal Behavior, Safety, Objectionable Content, Integrity and Authenticity, Respecting Intellectual Property, and Content-Related Requests and Decisions.[24] Some of these policies pertain to behaviors around which the matter of cultural relativism is not apparently relevant

[20] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 46 (2018).

[21] *Id.* at 52.

[22] Mark Zuckerberg, *Building Global Community*, https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/ (last visited Jan. 2, 2021).

[23] In 2018 and 2019, I served on an academic advisory group chaired by the faculty directors of the Justice Collaboratory of the Yale Law School called DTAG—Data Transparency Advisory Group. We released a report that assessed Facebook's methods of measuring and reporting on its Community Standards enforcement policies. The Just. Collaboratory, *Data Transparency Advisory Group*, https://law.yale.edu/justice-collaboratory/our-work/projects/data-transparency-advisory-group (last visited Jan. 2, 2021).

[24] Zuckerberg, *supra* note 22.

because the behavior at hand is not considered to vary along cultural lines. So, for example, Facebook's Community Standards lay out a policy against spam, which they describe as "content that is designed to deceive, or that attempts to mislead user to increase viewership" and that is designed "to artificially increase viewership or distribute content on masse for commercial gain."[25] All can agree that spam is a deceptive technique for distributing content, but what constitutes spam does not generate cross-cultural controversy. However, the broader Facebook policies on "authentic identity" do bring up fascinating cross-cultural issues on what anthropologists call "personhood," which, in the context of online social life, includes practices of anonymity and what it means to have multiple identities online.[26]

Many of the other areas addressed by Facebook's Community Standards clearly do pertain to norms that vary widely across the world. For example, Facebook regulates five kinds of "Objectionable Content": Hate Speech, Violent and Graphic Content, Adult Nudity and Sexual Activity, Sexual Solicitation, and Cruel and Insensitive. Attempts at monitoring and enforcing each of these areas has triggered many controversies, and Facebook and other platforms have had to frequently modify their policies in response to objections that, fittingly perhaps, have emerged on their own platforms (e.g., Instagram's "#freethenipple" hashtag).[27] For

[25] *Integrity and Authenticity*, FACEBOOK, https://www.facebook.com/communitystandards/integrity_authenticity (last visited, Jan 2, 2021).

[26] PAUL DOURISH & GENEVIEVE BELL, DIVINING A DIGITAL FUTURE: MESS AND MYTHOLOGY IN UBIQUITOUS COMPUTING 53 (2011)

[27] Julia Jacobs, *Will Instagram Ever 'Free the Nipple'?*, NEW YORK TIMES (Nov. 22, 2019), https://www.nytimes.com/2019/11/22/arts/design/instagram-free-the-nipple.html; Jillian York, *The Global Impact of Content Moderation*, ARTICLE 19 (Apr. 7, 2020), www.article19.org/resources/the-global-impact-of-content-moderation/; Frederik Stjernfelt & Anne Mette Lauritzen, *Nipples and the Digital Community*, in YOUR POST HAS BEEN REMOVED 95 (2020).

example, reference to cultural differences has been explicitly cited as being responsible for the difference between European and North American approaches to nudity and free speech, respectively. The argument is that Europeans are, for cultural reasons, more open to being exposed to nudity. In accordance with its absolutist values towards freedom of speech, the U.S. has been traditionally less willing to censor hate speech. This differs from countries like Germany, which have created freedom of speech restrictions in the context of Holocaust Denial.[28] Discussing cultural differences between the USA and France when it comes to internet governance, Jeffrey Rosen remarks, "Americans want to be famous, while the French want to be forgotten."[29] I will return to these cultural-based arguments in the next section, but for the moment, it is important to note the tension between "illusions of a borderless world" fostered by global internet platforms, and the realities of cultural borders.[30]

Applying human rights principles and institutions to the task of global content moderation represents an exercise that was not anticipated by the drafters of Universal Declaration of Human Rights (UDHR) and the subsequent human rights covenants. First, human rights law was created to check abuses of power by states, rather than private companies. Whereas human rights focus on the inherently universal concept of the "human" (as opposed to the citizen of a given nation-state), the equivalent concept in content moderation is the "user"—who elects to use what is often a nominally free service, one that admittedly has become like a utility.

[28] JEREMY WALDRON, THE HARM IN HATE SPEECH (2012); Noah Feldman, *Free Speech in Europe Isn't What Americans Think*, BLOOMBERG (Mar. 19, 2017, 10:33 AM), https://www.bloomberg.com/opinion/articles/2017-03-19/free-speech-in-europe-isn-t-what-americans-think; JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 145 (2019).

[29] Jeffrey Rosen, *The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google*, 80 FORDHAM L. REV. 1525 (2012).

[30] JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET?: ILLUSIONS OF A BORDERLESS WORLD (2006).

Another key difference is that a postulate of human rights is that they are supposed to be "fundamental," meaning that they pertain to "basic needs" and not "mere wants."[31] However, a glance at the community guidelines of platforms demonstrates that fundamental issues like the right to life and freedom of speech are present but so are matters that are hardly fundamental. Wanting to shield users from content that might make some uncomfortable is a matter of consumer satisfaction and in that sense represents part of the product that each platform is engineering. "Rights-oriented regulation" must make a case for stretching the concepts of human rights into these areas. [32]

But before discussing this issue, I want to turn to a discussion of some of the historical and ongoing concerns of the field of anthropology with the human rights movement that was launched with the Universal Declaration of Human Rights and the foundation of the United Nations at the end of World War II. My argument is that understanding the approach to culture and cultural relativism that has emerged in anthropology in the context of the conversations and debates about human rights give important insights on the challenges of internet governance, particularly as it pertains to content moderation.

FALSE UNIVERSALISM AND "THE RIGHTS OF MAN"

Finding ways to describe and translate the norms and standards by which people live but avoiding evaluating these based on one's own standards is historically a central puzzle of the anthropological endeavor. Anthropologists are trained to avoid "ethnocentrism" and to strive for an "emic approach"—to

[31] Burns Weston, *Human Rights*, ENCYCLOPEDIA BRITANNICA ONLINE, https://www.britannica.com/topic/human-rights (last visited Jan. 2, 2021).

[32] NICOLAS SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 9 (2019).

understand societies from an insider's perspective. Anthropologists naturally place cross-cultural scrutiny on attempts to create universal standards of any kind. In the case of the human rights movement, anthropologists have struggled with the concern that human rights standards reflect Western values and that they represent "false universalism."[33]

In 1947, Melville Herskovits, a leading anthropologist of the time, published a strongly worded rejection of a draft of the "Declaration on the Rights of Man," which was the document that would later become the Universal Declaration of Human Rights.[34] This document was being drafted in upstate New York by a commission organized by Eleanor Roosevelt. UNESCO had reached out to Herskovits for input as part of broad consultations with civil society.[35] In Herskovits's "Statement on Human Rights" he asserted that the document failed to address the following fundamental question: "How can the proposed Declaration be applicable to all human beings, and not be a statement of rights conceived only in terms of the values prevalent in the countries of Western Europe and America?"[36] Foreshadowing contemporary concerns regarding the universalist ambitions of "content moderation" and internet governance more broadly, Herskovits stated:

> Today the problem is complicated by the fact the Declaration must be of world-wide applicability. It must embrace and recognize the validity of many different ways of life. It will not be convincing to the Indonesian, the African, the Indian, the Chinese, if it lies on the same plane as like documents of an earlier period. The right of Man in the Twentieth Century

[33] Clive S Kessler, *Globalization: Another False Universalism?*, 21 THIRD WORLD Q. 931 (2000).
[34] Goodale, *supra* note 12, at 20-21.
[35] MARY ANN GLENDON, A WORLD MADE NEW: ELEANOR ROOSEVELT AND THE UNIVERSAL DECLARATION OF HUMAN RIGHTS (2001).
[36] The Executive Board of the Am. Anthropological Ass'n, *Statement on Human Rights*, 49 AMERICAN ANTHROPOLOGIST, NEW SERIES 539 (1947).

> cannot be circumscribed by the standards of any single culture, or be dictated by the aspirations of any single people. Such a document will lead to frustration, not realization of the personalities of vast numbers of human beings.[37]

What "single culture" did Herskovits have in mind in 1947 as he responded to a foundational document of the emergent United Nations that was being formed with the leadership of the United States and its triumphant allies in the aftermath of World War II?

Herskovits' objections to the nascent UDHR went beyond a concern about general ethnocentrism—an ethnocentrism that does not distinguish between who is doing the centering. Rather, he specifically warned against a kind of ethnocentrism that was combined with geopolitical power. He was particularly concerned about the newly unrivaled geopolitical power of the United States and its allies, notwithstanding their triumphant defeat of fascism. He identified the Western practice of "ascribing cultural inferiority" to non-Westerners as a key ideological buttress to a Western hegemony that he feared would not be overcome by a newly reconfigured geopolitical order that included the United Nations and a budding human rights system. He wrote:

> Doctrines of the "white man's burden" have been employed to implement economic exploitation and to deny the right of control their own affairs to millions of peoples over the world, where the expansion of Europe and America has not meant the literal extermination of whole populations. Rationalized in terms of ascribing cultural inferiority to these peoples, or in conceptions of their backwardness in development of their "primitive mentality," that justified their being held in the tutelage of their superiors, the history of the expansion of the western world has been marked by demoralization of the human personality and the

[37] *Id.* at 543.

> disintegration of human rights among the peoples over whom hegemony is established.[38]

From a contemporary perspective Herskovits' defense of cultural relativism manifested glaring weaknesses and contradictions.[39] Once more his "Statement on Human Rights" was not in itself an influential document among anthropologists moving forward, much less the drafters of the UHDR. For my purposes, what is noteworthy about this history is that Herskovits's is an early expression of the ways in which many anthropologists since that time have struggled to reconcile been concerned about how seemingly well-intentioned universalist principles can be exploitative. Mark Goodale argues that "…Herskovits drew from history in making the argument that declarations of human rights were often legal smokescreens for the oppression of one group of humans by another."[40]

Data Colonialism and Legal Smokescreens

In the context of the contemporary internet and the debates over how to govern it, scholars have identified troubling parallels between traditional colonialism and "data colonialism" as, according to Couldry and Mejias, "historic appropriation of land, bodies, and natural resources is mirrored today in this new era of pervasive datafication."[41] When it comes to the particular area of content moderation, "the specter of imperialism" is manifest as free speech policies generated in, for example, Silicon Valley and subsequently applied to the rest of the world.[42] We should not forget,

[38] *Id.* at 541.

[39] Alison Renteln, *Relativism and the Search for Human Rights*, 90 AM. ANTHROPOLOGIST 56, 67 (1988).

[40] Goodale, *supra* note 12, at 28.

[41] NICK COULDRY & ULISES A. MEJIAS, THE COSTS OF CONNECTION: HOW DATA IS COLONIZING HUMAN LIFE AND APPROPRIATING IT FOR CAPITALISM (2019).

[42] Goodale, *supra* note 12, at 64

of course, that this kind of intellectual inequality in which legal standards are manufactured in the first world and then exported to the third world is accompanied by traditional forms of labor inequality of the kind that media and technology scholar Sarah Roberts has documented. Roberts chronicles the ways in which the actual human labor of content moderation is exported to the third world in exploitative ways.[43]

In one version of an anti-imperialist critique of content moderation, the problem is not so much their conceptual ethnocentrism but rather their irresponsibility. In other words, platforms may fail to plan to account for the fact that they are generated in the world's metropoles but are put to the test in places where democracy is most fragile.[44] Legal scholar Michael Karanicolas recognizes "…the tension between implementing a moderation system which governs political discourse all over the world, but is disproportionately focused on impacts in the U.S."[45] He writes:

> This is always going to be a difficult balance to set, but it's made vastly harder by the differences across local contexts that are subject to the platforms' content moderation systems. A racially charged statement in Canada might cause psychological harm, but in Sri Lanka, it might lead to lynchings and communal violence. As recently as August, violent clashes in Bengaluru, India, were triggered by a Facebook post about the Prophet Muhammad. The potential harms, in other words, vary enormously.[46]

[43] SARAH ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).
[44] Karanicolas, *supra* note 16, at 183.
[45] Michael Karanicolas, *Moderate Globally Impact Locally: The Countries Where Democracy Is Most Fragile Are Test Subjects for Platforms' Content Moderation Policies* (Nov. 30, 2020), https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-countries-where-democracy-most-fragile-are-test-subjects-platforms.
[46] *Id*.

Here cultural difference is invoked as a way of acknowledging that different kinds of harm that may result from decisions that are made regarding permissible speech. In another version of an imperialist critique, the problem is that the United States has a particular 1st Amendment-based free speech tradition that, while it may or may not be appropriate for the United States, is a hazardous international export.[47] The hitch is that U.S.-based social media platforms export this approach while, arguably at least, "American approaches to freedom of expression diverge dramatically from those accepted in most of the remainder of the open and democratic world."[48]

Given the controversy around the globalization of content moderation standards, it is not surprising that people outside of the United States make nationalistic appeals to the defense of national sovereignty vis-a-vis other platforms and, at times, other standards. For example, such a scenario emerged in the aftermath of the call by the U.S. right wing to boycott Twitter and Facebook and enroll in Parler. In response to the alleged anti-conservative bias of Facebook and Twitter, Indians were presented with a homegrown alternative to U.S.-based social media called Tooter. Twitter and Facebook's content moderation policies have both run afoul of the Indian government in recent times.[49] In a recent case, the ruling the Bharatiya Janata Party criticized Twitter for allowing postings by a

[47] Cara Curtis, *Facebook's Global Content Moderation Fails to Account for Regional Sensibilities*, THE NEXT WEB, (Feb. 26, 2019), https://thenextweb.com/socialmedia/2019/02/26/facebooks-global-content-moderation-fails-to-account-for-regional-sensibilities/.

[48] Frederick Schauer, *Exceptional First Amendment*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS 48 (Michael Ignatieff ed., 2005)

[49] Chinmayi Arun, *Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms*, BERKMAN KLEIN CENTER FOR INTERNET AND SOC. (Jan. 23, 2018), https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32.

comedian that "lampooned India's Supreme Court" in ways that were deemed "obscene" and "degrading."[50] Tooter, the fledgling alternative micro-blogging platform, pitches itself as a *swadeshi* (Hindi for "native") version of Twitter that mirrors the aesthetics and format of Twitter.[51] The founders of Tooter are reported to have provided a nationalistic justification for their entry into the market: "We believe that India should have a Swadeshi social network. Without one we are just a digital colony of the American Twitter India Company, no different than what we were under the British East India Company."[52] The press coverage of Tooter in a lighthearted vein covered the memes that responded to the growth of Tooter. In a humorous way, most of the memes self-deprecatingly played with the idea that Tooter was a cheap Indian imitation of an American social media goliath. This expressed a dynamic that is well known across the developing world—wanting to value one's own while recognizing that one's own does not always measure up to global standards.[53] Tooter's creators, notwithstanding their anti-colonial pronouncements, explicitly promoted adherence to the First Amendment of the U.S. Constitution. They stated that their content

[50] Garavi Gujarat, *Twitter faces renewed heat in Indian over inaction against anti-court posts*, GG2, (Nov. 20, 2020), https://www.gg2.net/twitter-faces-renewed-heat-in-india-over-inaction-against-anti-court-posts/.

[51] Pallavi Punder, *What It's Like Using Indian, Twitter, Called Tooter*, VICE NEWS (Nov. 26, 2020), https://www.vice.com/en/article/88ax85/india-twitter-tooter-hindu-nationalism-alt-right-socil-media.

[52] Krishna Priya Pallavi, *Tooter, the Indian Twitter, sparks meme fest online*, INDIA TODAY (Nov. 26, 2020), https://www.indiatoday.in/trending-news/story/tooter-the-indian-twitter-sparks-meme-fest-online-best-reactions-1744229-2020-11-26.

[53] Brent Luvaas describes practice such as these in which graphic artists outside of the metropole take "cut and pasted" images from global marketing campaigns and repurposes them aesthetically in a subversive aesthetic process that he calls "brand vandalism." Brent Luvaas, *Designer Vandalism: Indonesia Indie Fashion and the Cultural Practice of Cut 'n' Paste*, 26 VISUAL ANTHROPOLOGY REV 1 (2010).

moderation policies would "not punish users for exercising their God-given right to speak freely."[54]

The global imposition of U.S-based legal standards via private media companies is one fear and one kind of legal smokescreen. However, the concern that scholars of the relationship between social media and global democracy express has to do with the ways in which government may appeal to internet sovereignty in order to justify the restriction of legitimate speech and opposition politics.[55] Here, governments do not protect themselves from outside impositions but rather "prevent data from flowing out through data localization" with authoritarian intent.[56] The metaphor changes from "keeping out" to "keeping in." In attempts to find a way to impose territorial models for controlling the flow of information, the matter of the nationality of the servers on which data will be stored becomes a subject of legislation and negotiation.

When it comes to how social media has been co-opted by authoritarian governments, Bradshaw and Howard describe how "computational propaganda is being used as a tool of information control in three distinct ways: to suppress fundamental human rights, discredit political opponents, and drown out dissenting opinions."[57] Beyond monitoring the application of content moderation standards by social media companies, many NGOs, think tanks, independent scholars, UN human rights mechanisms, and global internet watchdogs have emerged in recent years, intent on tracking the record of national governments who use the internet

[54] Pallavi, *supra* note 52.

[55] For a wide-ranging set of essays about these issues, see SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM (2020).

[56] Anupam Chander & Uyên P. Lê, *Data Nationalism*, 64 EMORY L. J. 677 (2015).

[57] Samantha Bradshaw & Philip Howard, *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*, PROJECT ON COMPUTATIONAL PROPAGANDA (2019).

to subvert democracy.[58] Cases in which national sovereignty is used to subvert democracy in a post-colonial context inevitably challenge relativistic, and anti-colonialist, concerns with Western cultural hegemony. This is a part of the puzzle that social media governance, as it emerges from the metropole, must solve without repeating the mistakes of previous efforts.

Geopolitics and the Internet: Revisiting the Asian Values Debate

Delving into the details of the ways in which social media is used for illiberal ends is beyond the scope of this essay. My objective here is to recognize a connection between newer controversies regarding internet governance and older controversies regarding general human rights-based approaches to international law. These controversies have been a chronic stumbling block in the development of the human rights movement, and indeed, the

[58] THE ELECTRONIC FRONTIER FOUND., http://www.eff.org (last visited Jan. 2, 2021); ARTICLE 19, http://www.article19.org (last visited Jan. 2, 2021); RANKING DIGITAL RIGHTS, http://rankingdigitalrights.org (last visited Jan. 2, 2021); GLOBAL NETWORK INITIATIVE, http://www.globalnetworkinitiative.org (last visited Jan. 2, 2021); David Morarand Bruna Martins dos Santos, *Online Content Moderation Lessons from Outside the US*, BROOKINGS (June 17, 2020), http://www.brookings.edu/blog/techtank/2020/06/17/online-content-moderation-lessons-from-outside-the-u-s/; ACCESS DENIED: THE PRACTICE AND POLICY OF GLOBAL INTERNET FILTERING (2008); ACCESS CONTESTED : SECURITY, IDENTITY, AND RESISTANCE IN ASIAN CYBERSPACE INFORMATION REVOLUTION AND GLOBAL POLITICS (Ronald Deibert et al. eds., 2012); Library of Congress, *Initiatives to Counter Fake News in Selected Countries* (2019), https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf; Collaboration on International ICT Policy for East and Southern Africa (CIPESA), *Despots and Disruptions: Five Dimensions of Internet Shutdowns in Africa* (2019)m https://cipesa.org/?wpfb_dl=2832020; Global Information Society Watch, *National and Regional Governance Forum Initiatives (NRIs)*, ASSOCIATION FOR PROGRESSIVE COMMUNICATIONS (2017), https://www.giswatch.org/sites/default/files/giswatch17_web.pdf; Ben Hassine, *Digital rights advocacy in the Arab world and the Universal Periodic Review*, ASSOCIATION FOR PROGRESSIVE COMMUNICATIONS (2016), https://www.apc.org/en/pubs/digital-rights-advocacy-arab-world-and-universal-p; David Kaye, *Report on Content Regulation*, Presentation to the 38th session of the Human Rights Council (Apr. 6, 2018). https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx.

development of the United Nations. Specifically, I am referring to the clash between the prerogative of national governments to exercise their sovereignty according to homegrown principles versus the global ambitions of reducing human misery by limiting abuses of power by the governments of the world. In the context of the universalism vs. human rights debate, one of the ways that this issue has manifested itself is in the form of the so-called "culture defense."[59]

In a narrower legal sense, the culture defense has to do with the admissibility of cultural evidence in the courtroom. However, in the broader context of human rights policy and law, the culture defense has to do with areas in which culturally rooted norms appear to be at odds with the law. In the most sensationalistic and inflammatory framing of the issue, what happens "when culture kills"? In her book on the subject, legal scholar Alison Dundes Renteln addresses classic cross-cultural cases that involve homicide, children, drugs, animals, marriage, attire and "The Dead."[60] How can one establish a universal age that divides childhood and adulthood? In what cases are the sacrifice of animals in religious contexts exempt from being treated as animal cruelty? In what context should polygynous marriages be tolerated and/or promoted? Renteln makes a case for the formal recognition in law of the culture defense but often scholars point to the cynical and self-serving ways in which governments invoke the culture defense to create a legal smokescreen for their abuses. This historic human rights dynamic resonates with contemporary concerns about internet governance.

One place where these lines of debate are well worn in general human rights discourse is in the so-called "Asian Values

[59] ALISON DUNDES RENTELN, THE CULTURE DEFENSE (2005).
[60] *Id*.

Debate."[61] The essence of the debate when it comes to human rights is whether or not the supposedly universal values that are expressed in the UDHR (or more broadly the European Enlightenment political philosophies that were drawn upon to construct the liberal democracies of the West) are compatible with "Asian culture."[62] Different authors populate the list of supposed cultural differences between East and West but the following traits are commonly cited: 1) emphasis on the community and societal harmony rather than individual personal fulfillment, 2) a sense of loyalty and duty toward one's family, 3) self-reliance and thrift, 4) a general tolerance of benign authoritarianism, 5) a stress on education, 6) respect for the elderly, 7) and respect for the accumulation of wealth.[63]

It would be easy to dismiss these characterizations as harmless generalizations, misguided orientalism and/or understandable expressions of regional pride in a post-colonial context, but they have geopolitical implications. Notwithstanding the 2012 ASEAN Rights Declaration, Asia does not have a fully developed regional human rights system. This stands in contrast to other world regions that do have regional instruments: the Organization of American States, the European Union, African Union, and the Arab League.[64] Whether or not the absence of inter-governmental human rights infrastructure in Asia has a significant

[61] Michael Freeman, *Human Rights and Real Cultures: Towards a Dialogue on 'Asian Values'*, 16 NETH. HUM. RTS. Q. 25 (1998). *See* Goodale, *supra* note 12, at 51-56 for an anthropologist's perspective on this polemic.

[62] Burns Weston, *The Universality of Human Rights in a Multicultured World: Toward Respectful Decision-Making, in* HUMAN RIGHTS IN THE WORLD COMMUNITY: ISSUES AND ACTION (2006); Adamantia Pollis & Peter Schwab, *Human Rights: A Western Construct with Limited Applicability in Human Rights*, HUMAN RIGHTS: CULTURAL AND IDEOLOGICAL PERSPECTIVES (1979).

[63] Harriet Samuels, *Hong Kong on Women, Asian Values, and the Law*, 21 HUM. RTS. Q. 707 (1999).

[64] Dinah Shelton, *Breakthroughs, Burdens and Backlash: What Future for Regional Human Rights Systems, in* HUMAN RIGHTS IN THE WORLD COMMUNITY: ISSUES AND ACTION (2006).

impact on the ground, what is important here is to note the particular ways in which cultural reasons have been invoked to justify the restriction of civil liberties and political freedoms in Asia.

At different moments in time, many government leaders in Asia have actively embraced Asian Values discourse, often in order to justify a perceived tradeoff between civil liberties and economic growth.[65] The late Lee Kuan Yew, the "founding father of Singapore," actively promoted Asian Values rhetoric and campaigns during his years as Prime Minister from 1959 to 1990.[66] In Neil Englehardt's article on the "Singaporean Confucian Ethics Campaign" of the 1980s, he demonstrates how Yew imposed a version of "Asian Values" on the Singaporean people in order to justify repressive policies.

These included measures such as "a restrictive press law designed to prevent criticism of the government, hampering freedom of expression and restricting access to alternative sources of information."[67] Englehardt describes how the campaign used the affinity that Singaporeans were inclined to have for elements of Chinese culture (from which in many ways they felt alienated) in order to promote values of obedience to authority and the "submergence of individual identity in collective identity."[68] Annette Marfording makes a similar critique of the ways in which the Japanese government and corporations cynically have enlisted the "Nihonjinron" literature (a genre that represents a Japanese take

[65] Han Sung-Joo, *Asian Values: An Asset or a Liability*, *in* CHANGING VALUES IN ASIA; THEIR IMPACT ON GOVERNANCE AND DEVELOPMENT (1999).

[66] Michael Barr, *Lee Kuan Yew and the 'Asian Values' Debate*, 24 ASIAN STUDIES REV. 309 (2000).

[67] Neil Englehart, *Rights and Culture in the Asian Values Argument: The Rise and Fall of Confucian Ethics in Singapore*, 22 HUM. RTS. Q. 548 (2000)

[68] *Id.* at 549

on Japanese culture) for their own advantage.[69] This history is important to remember as we consider contemporary geopolitics of the internet that are replaying themselves along the cold war lines, among other lines of contention, between the United States, Russia, and China,[70] the two other main social media platform-producing countries. The U.S. has accused them of extending authoritarian governance into the realm of the World Wide Web while the Snowden documents remind us of the use of the internet in mass surveillance by the United States.[71] If a truly global and just approach to social media governance is to emerge it will need to confront the claim that it must accommodate different cultures of privacy, surveillance, and conceptions of liberty.

CAPACITIES AND CAPABILITIES

How does a discipline that prides itself on the celebration of cultural difference and anti-imperialism[72] reconcile this fundamental commitment with seemingly misplaced appeals to culture and sovereignty that are used to justify the exercise of power by global elites at home and abroad—as illustrated by our brief description of the culture defense and the Asian Values debate? In other words, how can one separate genuine from spurious

[69] Annette Marfording, *Cultural Relativism and the Construction of Culture: An Examination of Japan*, 19 HUM. RTS. Q. 431 (1997).

[70] To cite just one example of saber rattling over internet espionage, the U.S. Secretary of State states on its website for "The Clean Network" government cybersecurity initiative, "We will keep doing all we can to keep our critical data and our networks safe from the Chinese Communist Party." *The Clean Network*, U.S. STATE DEPARTMENT, https://www.state.gov/the-clean-network/.

[71] Jeffrey Knockel et al., *We Chat, They Watch: How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus* (Citizen Lab Research Report No. 127, 2020) https://tspace.library.utoronto.ca/bitstream/1807/101395/1/Report%23127--wechattheywatch-web.pdf; TAYLOR OWEN & EMILY BELL, JOURNALISM AFTER SNOWDEN: THE FUTURE OF THE FREE PRESS IN THE SURVEILLANCE STATE (2017).

[72] Peter Pels, *What has anthropology learned from the anthropology of colonialism?*, 16 SOC. ANTHROPOLOGY 280 (2008).

representations of culture?[73] How can one celebrate cultural difference while also recognizing that appeals to culture can be used to oppress? And how has the discourse of human rights provided intellectual leverage with which to resolve these related dilemmas? Whereas some have been willing to declare human rights universalism provisionally victorious,[74] anthropologists have found ways to defend the concept of culture, a focus on the human and specific versions of cultural relativism while embracing the relativistic spirit of the American Anthropological Association's (AAA) original dissent.[75] Mark Goodale captures this paradox when he states, "…what human rights needs is more humanist restraint and appreciation for particularity and less enlightenment triumphalism."[76] Anthropological supporters of the human rights movement lend their support by resisting the temptation to recognize the victory of human rights universalism. This is what Marie-Benedicte Dembour means when she describes the "pendulum" that anthropologists walk between relativism and universalism in which they "err uncomfortably between the two poles."[77]

In 1999 the membership of the American Anthropological Association (AAA) adopted a statement on human rights that represented a formal reversal from the contrary stance penned by Herskovits on behalf of the AAA in 1947. One of the ways that this

[73] Richard Handler & Jocelyn Linnekin, *Tradition: Genuine or Spurious*, 97 J. AM. FOLKLORE 273 (1984).

[74] Jack Donnelly's work is known for the strongest and most celebratory defense of universalism; JACK DONNELLY, INTERNATIONAL HUMAN RIGHTS (2nd ed., 1999); JACK DONNELLY, UNIVERSAL HUMAN RIGHTS IN THEORY AND PRACTICE (2003).

[75] Engle, *supra* note 13.

[76] Goodale, *supra* note 12, at 16.

[77] Marie-Benedicte Dembour, *Following the Movement of the Pendulum: Between Universalism and Relativism*, *in* CULTURE AND RIGHTS: ANTHROPOLOGICAL PERSPECTIVES 59 (Jane Cowan et al. eds. 2001).

document juggles the paradox that I mention above is that rather than defending any particular set of rights it defends the "capacity for culture." The 1999 Statement on Human Rights states:

> The capacity for culture is tantamount to the capacity for humanity. Culture is the precondition for the realization of this capacity by individuals, and in turn depends on the cooperative efforts of individuals for its creation and reproduction. Anthropology's cumulative knowledge of human cultures, and of human mental and physical capacities across all populations, types, and social groups, attests to the universality of the human capacity for culture. This knowledge entails an ethical commitment to the equal opportunity of all cultures, societies, and persons to realize this capacity in their cultural identities and social lives. However, the global environment is fraught with violence which is perpetrated by states and their representatives, corporations, and other actors. That violence limits the humanity of individuals and collectives.

Though 50 years earlier Herskovits had rejected the UDHR on the grounds of its Eurocentrism, the 1999 Statement endorses the UDHR (and subsequent UN Human Rights Conventions) as tentative "working definitions" of "respect for concrete human differences." It reminds us that these UN formulations of human rights represent only "the abstract legal uniformity of the Western tradition." The statement presents the definition of human rights as a "constantly evolving" process and invites members of the AAA to get "involved in the debate on enlarging and understanding human rights on the basis of anthropological knowledge."[78]

This tentative embrace of human rights via the notion of a "capacity for culture" parallels the "capabilities approach" to human rights that political philosopher Martha Nussbaum and development

[78] Committee for Human Rights, *1999 Statement on Human Rights*, AM. ANTHROPOLOGICAL ASS'N, http://humanrights.americananthro.org/1999-statement-on-human-rights/ (adopted by the AAA membership, June 1999).

economist Amartya have developed in a much more elaborate and programmatic way.[79] The capabilities approach has been applied in a wide variety of ways, but as far as the matter of the universality of human rights is concerned, it adds philosophical heft to documents like the UDHR that might otherwise be viewed as sterile laundry lists of rights. Nussbaum takes the rights of the UDHR (e.g., Article 3 on the right to life and Article 19 on the right to freedom of expression) and shows how they correspond with essential capabilities that all human beings share.[80] For example, Nussbaum takes Article 18 of the UDHR ("freedom of thought, conscience and religion") and notes these are expressed as basic entitlements. She then generates a list of the underlying "capabilities" that correspond to each of the human rights in the UDHR. In the case of Article 18 the capability that corresponds to this article is "practical reason," which she defines in the following way: "Being able to form a conception of the good and to engage in critical reflection about the planning of one's life. This entails protection for the liberty of conscience and religious observance."[81]

The following two aspects of her approach are the most relevant in the context of the intersection of human rights, anthropology and internet governance.

First, Nussbaum grounds her approach in a kind of universalism that aspires to not be grounded in any particular articulation of human rights nor in any particular cultural tradition. Rather she grounds them in the universality of the human person and the fundamental capabilities ("life," "bodily health," "bodily

[79] MARTHA NUSSBAUM, CREATING CAPABILITIES: THE HUMAN DEVELOPMENT APPROACH (2011); AMARTYA SEN, COMMODITIES AND CAPABILITIES (1999).
[80] Martha Nussbaum, *Capabilities, Human Rights and the Universal Declaration*, *in* THE FUTURE OF INTERNATIONAL HUMAN RIGHTS (1999).
[81] *Id.*

integrity," "senses," "emotions," "practical reason," "affiliation," "friendship," "play," etc.) that we all share regardless of how our cultures shape their expression. Addressing a fundamental anthropological ambition, she creates a framework that establishes "the unity of humankind" as a point of departure. Secondly, the capabilities approach creates a human rights methodology that is based on "appreciation for particularity" (to return to the above quotation from Mark Goodale).

This sets the stage for an approach to human rights that is referred to as the "indivisibility" of human rights and "human rights holism."[82] A holistic approach requires us to consider the ways in which the interaction between human rights enables their full enjoyment. Particularly in a polarized cold war context where the Socialist and Non-Aligned countries argued that social and economic rights were more fundamental than the civil and politics that were prioritized by the Liberal Democracies, the capabilities approach refuses to create a "hierarchy of rights" by insisting on drawing our attention to their relationship.[83] This perspective is most succinctly captured by Amartya Sen's famous thesis: "No famine has ever taken place in the history of the world in a functioning democracy."[84]

What does the capabilities approach and the appeal of it to anthropologists[85] have to do with the issue of social media governance? The capabilities approach is a methodology that is about making judgments on whether a person is suffering harm, and it requires us to dig into the details of that person's life as a member

[82] A. Belden Fields, *A Holistic Approach to Human Rights*, *in* RETHINKING HUMAN RIGHTS FOR THE NEW MILLENNIUM (2003).
[83] Tom Farer, *The Hierarchy of Rights*, 8 AM. U. INT'L L. REV. 115 (1992).
[84] AMARTYA SEN, DEVELOPMENT AS FREEDOM 16 (1999).
[85] Mark Goodale, *Introduction: Human Rights and Anthropology*, in GOODALE, *supra* note 12.

of their social and political worlds in order to make these kinds of determinations.[86] It is about making cross-cultural determinations. It requires the adjudicator to go beyond, “How satisfied is person A?” and ask, “What is A actually able to do and be?”[87] What are their ambitions and what are the opportunities that are available to that person? In the words of Nussbaum, “It looks at not what people feel about what they do, but about what they are actually able to do.”[88]

So, for example, if we are to determine whether a woman is enjoying the right to vote we must also ask whether her mobility and access to education and employment are not limited by political and/or cultural restrictions. She may have the formal right to vote but in the context of her particular life circumstances we may determine that she does not truly enjoy that right. Remedies would also need to avoid narrowing in on the formalities of the voting system and address broader considerations such as gender discrimination in the areas of healthcare, education, transportation, etc. Anthropological methodologies, such as participant-observation and other ethnography, provide the in-depth understandings of people’s “everyday life” that are required to put the capabilities approach into practice.

Mental Autonomy and Architectural Regulation

What might this rights-oriented and ethnographic approach to making determinations look like in the emergent context of content moderation and internet governance? Legal scholar Eveyln Aswad has detailed one such approach to regulating privacy, censorship and free speech on the internet that is grounded in appeal

[86] Samuel Martinez, *Searching for a Middle Path: Rights, Capabilities, and Political Culture in the Study of Female Genital Cutting*, 22 THE AHFAD J. 31 (2005).
[87] Nussbaum, *supra* note 80.
[88] *Id*.

to Article 19 of the ICCPR concerning the right to "hold opinions without interference."[89] Her perspective builds on the institutional efforts in this area within the UN human rights system including the efforts of David Kaye and Irene Khan, the former and current Special Rapporteur on the Right to Freedom of Opinion and Expression.[90] Aswad argues that the wording of Article 19 invites us to think more expansively and holistically about what it means to enjoy the freedom of speech in an internet age characterized by the proliferation of a digital economy that runs on "digital extraction and the monetization of digital data."[91] The threats to the enjoyment of human rights in her opinion stem from the following aspects of the internet companies: (1) designing digital products to maximize time spent on platforms, (2) leveraging user engagement to continuously extract personal data, and (3) using and selling that data to target users with highly particularized information in order to affect their views and behavior.[92]

Rather than focusing on whether any particular kind of speech should or should not be allowed on a platform, Aswad's approach asks us to dig deeper and ask whether these aspects of the "business model" of the platforms infringes on the "basic ability to think and form opinions."[93] In her approach, a human rights based approach to platform governance must be dedicated to protecting the "mental autonomy" of the public that at present is at a "high risk of manipulation."[94] She concludes by offering a series of

[89] Evelyn Aswad, *Losing the Freedom to Be Human*, 52 COLUM. HUM. RTS. L. REV. 306 (2020).
[90] *Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN OHCHR, https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/mandate.aspx (last visited Jan. 7, 2020).
[91] Aswad, *supra* note 89, at 369.
[92] *Id.* at 369.
[93] *Id.* at 310.
[94] *Id.* at 369.

recommendations for leveraging Article 19 of the ICCPR and the UN Guiding Principles on Businesss and Human Rights (UNGPs), such as regulations related to Free, Prior and Informed Consent and the "deployment of digital literacy campaigns."[95]

Aswad does not mention the capabilities approach nor the importance of ethnographic research, but her work obviously resonates with a human-rights oriented anthropology. It is built on the recognition of fundamental human abilities and an appreciation of the fact that these cannot be understood in isolation from the broader constraints and possibilities that exist in a person's life. Laura DeNardis and Francesca Musiani have written about the "turn to infrastructure"[96] in internet governance that brings our attention to the ways in which governing effectively in the internet world requires attention to the ways in which control is embedded in the structures of the platforms. They caution us on approaches that pay too much attention to just "content and expressive freedom."[97] Laura DeNardis writes:

> …the diffusion of digital technologies into the material world necessitates a radical reconceptualization of freedom and human rights. Traditional notions of Internet freedom are disconnected from actual technical, political, and market conditions. "Internet freedom" usually pertains to content, especially freedom of expression, intellectual property rights, and freedom from government regulation of content. Rarely has it involved technical architecture itself, although interestingly the philosophical principles of freedom and openness have some historical roots in the Internet's engineering design community. When human rights concerns do invoke infrastructure, this

[95] *Id.* at 368.

[96] Laura DeNardis & Francesca Musiani, *Governance By Infrastructure*, *in* THE TURN TO INFRASTRUCTURE IN INTERNET GOVERNANCE (Francesca Musiani et al. eds. 2016).

[97] LAURA DENARDIS, THE INTERNET IN EVERYTHING: FREEDOM AND SECURITY IN A WORLD WITH NO OFF SWITCH 183 (2020).

> connection has primarily focused on access rights that affect the flow of content, such as broadband penetration rates or net neutrality, both infrastructure issues that reside very close to human users rather than embedded in technical architecture."[98]

Lee Tien makes similar observations regarding "architectural regulation" in which control mechanisms are "embedded into settings of equipment." [99] His critique of this kind of hidden regulation in which "code is law,"[100] as opposed to traditional "sanctioned-backed" legal approaches, resonates in stimulating ways with the capabilities approach to human rights. Architectural approaches focus on the mechanisms through which computer and network infrastructure limit and channel behavior in often unseen ways as they constrain even the ability to imagine other choices and possibilities. Considering these dynamics is critical if we are to productively apply human rights principles to the particular challenges of internet governance. Sen and Nussbaum, engaging different literatures and contexts, have provided invaluable insights into holistic and cross-cultural ways of doing this.

Vernacularization and Translation

The late Sally Engle Merry was a leading anthropologist who wrote about the internationalization of the human rights movement, particularly regarding the worldwide struggle against domestic violence against women.[101] Her work on the subject provides examples of how the field can both dedicate itself to

[98] *Id.* at 164-65.

[99] Lee Tien, *Architectural Regulation and The Evolution of Social Norms*, 7 YALE J.L. & TECH. 1 (2005).

[100] Lawrence Lessig, *Code is Law: On Liberty in Cyberspace*, HARV. MAG. (Jan.-Feb. 2000).

[101] Tragically, Sally Engle Merry died on September 8, 2020. Philip Alston et al., *In Memoriam: Sally Engle Merry*, CHR&GJ (Sept. 9, 2020) http://chrgj.org/2020/09/09/in-memoriam-sally-engle-merry/.

understanding culture-transforming global social movements while also affirming its traditional commitment to cultural relativism. She studied the ways in which human rights discourses traveled—from global campaigns to transnational activists to local practitioners and then back again. She saw this not as a process of imposition nor the replacement of one culture by another but rather as a process of "translation" which she called human rights "vernacularization."

Ideas like rights are said to be vernacularized when they "are adapted to local institutions and meanings."[102] Resisting the pressures of the universalism-relativism debate, Engle Merry paid attention to the ways in which human rights circulate between global contexts like UN conferences and academic panels to the places and institutions where programs to deter gender violence were being put into practice like India, China, Fiji, Hong Kong, Hawai'i, and Massachusetts.[103] Michael Ignatieff has recognized the importance of this process of vernacularization of human rights in the following way: "As a language of moral claims, human rights has gone global by going local, by establishing its universal appeal in local languages of dignity and freedom."[104] If human rights principles and institutions are going to be used effectively in content moderation policy than they must find a way to craft standards that are not so flexible that they are meaningless while they also must appeal to local concepts that have particular resonance in their respective contexts.

In her global study of the globalization of human rights-based programs against gender violence, Sally Engle Merry

[102] Sally Engle Merry, *Transnational Human Rights and Local Activism: Mapping the Middle*, 108 AM. ANTHROPOLOGIST 39 (2006).
[103] SALLY ENGLE MERRY, HUMAN RIGHTS & GENDER VIOLENCE: TRANSLATING INTERNATIONAL LAW INTO LOCAL JUSTICE (2006).
[104] Michael Ignatieff, *Introduction*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS 37 (Michael Ignatief ed., 2005).

observed that in many of the places that she visited there were cultural and political barriers from even recognizing gender violence as a serious social problem at all. For example, in India, "cruelty" had been the term that historically had been used to label what is now widely called "domestic violence."[105] As she studied the ways in which the new concept of domestic violence was mobilized in each of the places that she studied, she attempted to trace how human rights principles, such as those expressed in CEDAW (the UN Convention on the Elimination of All Forms of Discrimination Against Women), were adapted and transformed by cultural "translators" who bridged the global women's rights movement and the various local contexts.

In the case of India, she examined the cultural specificities of gender violence in a context in which the politics of dowry payments in marriages sometimes spiraled out of control. She noted that in the "criminalization" stage of Indian initiatives against domestic violence, various strategies were used that took these features into account. In the 1980s, special police stations were formed that were focused on dowry conflicts. In the 1990s, all-women police units and specialized family courts were formed. The ironic observation that she made was that initiatives in the area of domestic violence became more harmonized with international principles and practices rather than less harmonized. This was partially the result of pressure from transnational Indian women's rights activist who were guided by CEDAW.[106] In her multi-country comparison she revealed that "the most striking finding is the extent to which despite significant variation in cultural background, political power, and history of each country, the palette of reforms

[105] MERRY, *supra* note 103, at 139.
[106] *Id.* at 139-43.

is similar."[107] At the end of day, she confesses that much of the translation into local cultural terms is "a kind of window dressing."[108]

As an activist committed to global women's solidarity, she refused to accept culturally-rooted justifications for the violence that women suffered daily—whether that be in the India or the United States. On the other hand, she needed to recognize that the effectiveness of social movements that were perceived as alien sometimes faced obstacles to their acceptance but at other times benefitted from their foreignness. As an Anthropologist she might be inclined to at least hope for the possibility of acknowledging homegrown approaches to domestic violence that were built on a primarily pre-existing cultural substrate. But she found that, to the contrary, transnational domestic abuse intervention programs "acquire local symbolic elaboration, but retain their fundamental grounding in transnational human rights concepts of autonomy, individualism, and equality." In other words, they were "appropriated and translated but not fully indigenized."[109]

Merry's answer to the specter of "moral imperialism" involves two parts. The first is essentially an ethnographic response. She creates a framework for studying the very process into which people make the difficult tradeoff between pro-rights reform and the cultural transformations that accompany them. Rather than seeing global human rights reform movements as a purely political phenomenon she encourages us to view them as sites of transformation in which appeals to culture are made strategically to "vernacularize" and "indigenize" global human rights norms. She states this succinctly, "Instead of asking if human rights are a good

[107] *Id.* at 177.
[108] *Id.*
[109] *Id.* at 178.

idea, [an anthropological approach to human rights] explores what difference they make."[110] The second part of her response that I want to highlight in the context of this paper is the fact it is represents a quiet, but in its own way, quite forceful defense of human rights universalism. After all, why would a field that celebrates cultural difference and cultural sovereignty accept the homogenization that comes along with human rights reform?

When it comes to the reform of social media content moderation policies, what are the benefits of a consideration of Merry's approach to human rights vernacularization? The content moderation policies of all of the major platforms have fallen into the same trap. As we noted above, they have created a single set of standards and they use computers and reviewers to attempt to apply these standards to the online behavior of the people who use their services. They have understood this process as universalist exercise that requires them to be inflexible precisely because it is a universalist exercise. As Merry has shown us, however, the promotion of universalism does not require inflexibility. Rather it is an invitation for policy makers to vernacularize universal principles through careful consideration of the cultural milieus in which they will be designed and implemented as well as, more importantly perhaps, a careful consdiration of the ways in which the internet intersects with the daily lives of people across the globe.

CONCLUSION

It is beyond the scope of this paper to delve into the details of the virtues of a human rights approach to content moderation and internet governance. In recent books and other forums, scholars such as Tarleton Gillespie, Nicholas Suzor and David Kaye have made

[110] *Id.* at 39.

robust arguments in favor of such an approach, which Suzor calls "New Constitutionalism."[111] It is even more beyond the scope of this paper to delve into details about the virtues of a human rights approach to global governance in general.[112] This has been an exercise in triangulation in which I put scholarship on human rights governance in conversation with scholarship on internet governance in conversation with anthropological approaches to human rights. I argue that the ongoing work of reforming content moderation policies will benefit from understanding the histories and debates that I have outlined here in order to avoid some of the pitfalls that are inherent in this particular kind of global governance. The objective of this exercise has been to put into conversation concepts that are isolated from each other such as "capacity for culture," the "capabilities approach," "architectural regulation," "mental autonomy," and "vernacularization." These topics lie at the intersection of anthropology, political philosophy, and media studies.

What lessons should we take away from this exercise in triangulation? Anthropology has long struggled with a concern about whether the idea of human rights (or any other globalizing ideology, for that matter) is or will become a technique of "moral imperialism." I have briefly outlined one intellectual tradition within the discipline that has arrived at a version of human rights universalism—one that is composed, in the word of Mark Goodale, in a "minor key."[113] Goodale states that "….an anthropology of human rights envisions a future transnational or post national normative framework that is based on the imperatives of ethical

[111] *See* Gillespie, *supra* note 20; Kaye, *supra* note 16; Suzor, *supra* note 32.
[112] For a general discussion of the human rights concept in theory and practice, see HUMAN RIGHTS IN THE WORLD COMMUNITY: ISSUE AND ACTIONS, (Burns Weston & Anna Grear eds., 4th ed. 2016).
[113] Goodale, *supra* note 12, at 132.

restraint, humility, and legal pluralism."[114] Given the geopolitics of internet governance and social media content moderation, human rights principles represent a critical tool in establishing cross-cultural legitimacy for the new strategies of governance that will emerge. But with the unprecedented reach and power of global technologies of communication and control, the importance of truly universal solutions, ones that will be embraced across the globe, is unmistakable.

This paper is intended to point internet governance scholars in the direction of a body of literature in anthropology that might be overlooked and that provides an important set of questions and methodologies that are worthy of review and consideration. Although I have not proposed concrete examples of how future reforms of internet governance should look, I hope that this exercise gives us at least a better idea of how these reforms should sound.

[114] *Id.* at 133.

REIMAGINING SOCIAL MEDIA GOVERNANCE: HARM, ACCOUNTABILITY, AND REPAIR

Sarita Schoenebeck & Lindsay Blackwell[*]

INTRODUCTION

Social media companies have attracted widespread criticism for the proliferation of harmful behaviors on their platforms. Individual users levy hate speech and harassment at their peers; state actors manipulate networks of fraudulent accounts to propagate misinformation; extremist groups leverage recommendation systems to recruit new members. While these and similar harmful behaviors are extensions of existing social phenomena and not inventions of the internet age, they are exacerbated and intensified by the specific technological affordances of social media sites, including visible network relationships, quantified social endorsement (e.g., "likes" and follows), and algorithmic feeds designed to maximize social engagement.

Because of the scale at which contemporary social media platforms operate—Facebook recently reported 1.84 billion daily active users[1]—traditional forms of social media governance, such as the appointment of volunteer moderators, have struggled to keep apace. Social media companies have attempted to address these concerns by developing formal content moderation policies and enforcement procedures, but they are not made transparent to users,

[*] Sarita Schoenbeck, Professor, School of Information, University of Michigan; Lindsay Blackwell, PhD Candidate, School of Information, University of Michigan.

[1] *Fourth Quarter 2020 Results Conference Call*, FACEBOOK (Jan. 27, 2021), http://s21.q4cdn.com/399680738/files/doc_financials/2020/q4/FB-Q4-2020-Conference-Call-Transcript.pdf.

both in process and outcome.[2] Scaled content moderation also requires significant human labor—typically outsourced to third-party contractors who earn relatively low wages for work that is both physically and emotionally taxing[3]—to review individual pieces of content for potential policy violations, which results in delayed response times and backlogs of lower-priority violations.

Though regulators, researchers, and practitioners alike agree that change is needed, experts disagree on best paths forward. We propose a reframing of social media governance focused on repairing harm. Repairing harm requires recognizing that harm has occurred; centering the needs of individuals and communities who experience harm; and accepting accountability for the harm, both for the specific instance of harm and its root causes.

We first review prominent paradigms for the regulation of online behavior, from the 1980s through the early 2020s. Then, we discuss common categories of harm experienced on or created by social media platforms, including the consequences of inadequate platform governance. Drawing on principles of retributive, restorative, and transformative justice, we propose social media governance frameworks for better addressing those harms. We argue that, although punishment is sometimes necessary, a solely punitive model of governance is insufficient for encouraging compliance or for deterring future harm. We conclude with several

[2] *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, https://santaclaraprinciples.org (last visited Jan. 10, 2021) [hereinafter *The Santa Clara Principles*]; JILLIAN C. YORK, SILICON VALUES: THE FUTURE OF FREE SPEECH UNDER SURVEILLANCE CAPITALISm (2021); Ben Bradford et al., *Report Of The Facebook Data Transparency Advisory Group*, JUSTICE COLLABORATORY (2019), https://academyhealth.org/sites/default/files/facebookdatatransparencyadvisorygraoupreport52119.pdf; TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

[3] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

key shifts for transforming platform governance, focusing on the structural changes required to both repair and reduce harm.

Position Statement

Researchers are not separate from the social processes they study; our values, beliefs, and experiences inevitably influence our analyses. As such, it is not possible to appropriately position any work without first understanding the relative position of its authors. Both authors of the present work are cisgender women; one author is queer. One author is white, and the other is white-presenting; though we draw from foundational scholarship by a range of scholars to support our analyses, the absence of experiences from or interpretations by Black, Indigenous, and people of color is a significant limitation of this work. It is similarly limited in its cultural perspective, with both authors having lived, been educated, and been employed in the United States. Although one author's experiences of disability inform her perspective, disability justice is also out of scope for the present work. Finally, one author is an academic researcher and tenured professor at a research institution in the midwestern United States; the other is a student at this same institution and has worked as a corporate social media researcher for four years.[4] Both authors are social media users, have personally experienced online harassment, and have studied intersections between social media behavior and governance in both academia and industry.

[4] Blackwell has worked full-time at Facebook and Twitter. Schoenebeck has consulted with Twitter and received funding from Instagram, Facebook, Mozilla, and Google. This work was not directed by, nor does it express the opinions of, any company.

PARADIGMS OF SOCIAL MEDIA GOVERNANCE

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users, including but not limited to flaming (or the use of inflammatory language, name calling, or insults); doxing (or the public release of personally identifiable information, such as a home address or phone number); impersonation (or the use of another person's name or likeness without their consent); and public shaming (or the use of social media sites to humiliate a target or damage their reputation). While online harassment is sometimes depicted as an outlier or fringe behavior, an overwhelming number of social media users have experienced or witnessed some form of online harassment.[5] Harassment tactics are sometimes employed concurrently, particularly when many individuals, acting collectively, target just one individual (sometimes referred to as "dogpiling"). One individual may also harass another, as is often the case in instances of cyberbullying[6] and non-consensual intimate image sharing (also known as "revenge porn"), in which sexually explicit images or videos are distributed without their subject's consent, often by a former romantic partner.[7] Online harassment experiences can range from a single instance to repeated harassment over a sustained period of time; similarly, given the networked

[5] Maeve Duggan, *Online Harassment 2017*, PEW RESEARCH CENTER: INTERNET & TECHNOLOGY (Jul. 11, 2017), http://www.pewinternet.org/2017/07/11/online-harassment-2017/.

[6] Zahra Ashktorab & Jessica Vitak, *Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design With Teenagers*, *in* PROCEEDINGS OF THE 2016 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYS. 3895 (2016); Peter K. Smith et al., *Cyberbullying: Its Nature and Impact in Secondary School Pupils*, 49 J. CHILD PSYCH. & PSYCHIATRY 376 (2008).

[7] CARRIE GOLDBERG, NOBODY'S VICTIM: FIGHTING PSYCHOS, STALKERS, PERVS, AND TROLLS (2019); Danielle Keats Citron, *A New Compact for Sexual Privacy*, William & Mary L.R. (forthcoming), https://papers.ssrn.com/abstract=3633336 (last visited Dec. 7, 2020); Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345 (2014).

nature of social media platforms, targets may be harassed by one perpetrator or thousands. These attributes often overlap, especially in the case of coordinated, networked harassment campaigns that are long-term and large-scale.

Regulating behavior is complex, and contemporary social media platforms face numerous challenges. Some are challenges of scale: monolithic approaches to online governance approaches start to crumble at the scale of millions or even billions of diverse users.[8] Others are challenges of adaptability: best practices in one community or platform may fall short in another, particularly on large, global platforms where diverse individual and cultural norms intersect. They may also be failures of anticipation: few could have foreseen the concentration of global power now held by a handful of corporate leaders.

Social media governance is both social and technical; the sociotechnical perspective[9] describes how social and technical aspects of systems are necessarily interrelated and cannot be disentangled. In other words, we cannot design a technological system without also considering its social impacts, and we cannot understand the social impacts of a system without also considering its design and politics. A sociotechnical lens of social media governance argues that design principles and practices will inevitably shape how social behavior is governed online, and vice versa. This section establishes four major paradigms of social media governance: normative, distributed, algorithmic, and retributive

[8] GILLESPIE, *supra* note 2; ROBERTS, *supra* note 3.

[9] Mark S. Ackerman, *The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility*, 15 HUM.–COMPUT. INTERACTION 179 (2000).

regulation.[10] These paradigms are overlapping, both temporally and categorically, and reflect evolving social behaviors and technological affordances.

Normative Regulation

The earliest paradigm of governance, emerging in the 1980s[11], involved establishing and reinforcing norms for good behavior, sometimes assigning community members special privileges (e.g., administrator or moderator status) to enforce those norms.[12] This early paradigm also saw the introduction of specialized moderation tools to support regulation, such as reporting, flagging, and editorial rights.[13]

Online communities continue to rely on normative regulation today, both through formal rules—typically asserted by community guidelines and enforced via content moderation[14]—as well as through unstated, informal norms that are learned through

[10] An early version of these paradigms was developed in Lindsay Blackwell et al., *When Online Harassment is Perceived to be Justified*, *in* INTERNATIONAL AAA CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM) (2018).

[11] HOWARD RHEINGOLD, THE VIRTUAL COMMUNITY: HOMESTEADING ON THE ELECTRONIC FRONTIER (2000); JULIAN DIBBELL, MY TINY LIFE: CRIME AND PASSION IN A VIRTUAL WORLD (1998).

[12] Eshwar Chandrasekharan et al., *The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales*, 2 PROC. ACM HUM.-COMPUT. INTERACT. 32:1 (2018); DIBBELL, supra note 11; Robert Kraut & et al., *The HomeNet Field Trial of Residential Internet Services*, 39 Commc'n of the ACM 55 (1996); ROBERT E. KRAUT ET AL., BUILDING SUCCESSFUL ONLINE COMMUNITIES: EVIDENCE-BASED SOCIAL DESIGN (2012); Cliff Lampe & Paul Resnick, *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Spac*e, in PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 543 (2004).

[13] Lindsay Blackwell et al., *Classification and its Consequences for Online Harassment: Design Insights from Heart*Mob, 1 PROC. ACM HUM.-COMPUT. INTERACT. 19 (2017); J. Nathan Matias et al., *Reporting, Reviewing, and Responding to Harassment on Twitter* (2015), http://womenactionmedia.org/twitter-report; Jessica A. Pater et al., *Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms*, in PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON SUPPORTING GROUP WORK 369 (2016).

[14] ROBERTS, *supra* note 3.

participation in the community.[15] While social media companies have largely relied on prescriptive norms (i.e., explicit rules) to govern user behavior, descriptive norms—the implicit social expectations we learn by observing how others interact in a given space—are much more powerful at shaping behavior. Prescriptive norms establish how people *should* behave, descriptive norms describe how people are already behaving—creating what Cialdini describes as "a decisional shortcut" when other people are choosing how to behave.[16]

Although normative regulation allows communities to self-govern in ways that are aligned with their specific values and priorities, these strategies are more effective in communities with clearly-established boundaries, such as individual subreddits.[17] Many popular platforms, such as Twitter and TikTok, lack formal community infrastructures, which constrains their ability to rely on normative regulation. Even in online spaces with a clear sense of community, antisocial norms—for example, norms that encourage discrimination, hatred, racism, and other harms—may also emerge and can persist if left unchecked.[18]

Distributed Regulation

A second paradigm saw the rise of crowd-sourced approaches to behavioral regulation, first popularized by platforms

[15] J. Nathan Matias, *Preventing Harassment and Increasing Group Participation Through Social Norms in 2,190 Online Science Discussions*, 116 PNAS 9785 (2019); Chandrasekharan et al., supra note 12.

[16] Robert B. Cialdini, Carl A. Kallgren & Raymond R. Reno, *A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior*, 24 ADVANCES IN EXPERIMENTAL SOC. PSYCH. 201 (1991).

[17] Matias, *supra* note 15; Chandrasekharan et al., *supra* note 12.

[18] Eshwar Chandrasekharan et al., *You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech*, 1 PROC. ACM HUM.-COMPUT. INTERACT. 31:1 (2017); Kishonna L. Gray, *Black Gamers' Resistance*, *in* RACE AND MEDIA: CRITICAL APPROACHES 241 (Lori Kido Lopez ed., 2020).

in the early 2000s (e.g., Slashdot and Digg) and still in use by some contemporary platforms (e.g., Reddit and Wikipedia). This model of governance—what Grimmelmann characterizes as distributed moderation[19]—traditionally relies on scalable feedback mechanisms (e.g., upvotes and downvotes) to establish the appropriate enforcement action. For example, a post that receives a high volume of upvotes may be featured more prominently; conversely, a post receiving a high volume of downvotes may be a candidate for deletion.

Distributed and normative regulation overlap in their reliance on shared community norms to govern behavior. Thus, while crowd-sourced governance can be an effective mechanism for reducing harmful content, this is ultimately dependent on the specific values of a given community. Some communities may embrace offensive, violent, or other kinds of damaging content as desirable,[20], rendering distributive regulation effective at enforcing the community's values but not at discouraging harm. Distributed moderation is also vulnerable to manipulation; most technical feedback mechanisms are easily manipulated by smaller factions of users (e.g., recruiting additional users to artificially inflate vote counts), sometimes with the express purpose of amplifying harm.

Algorithmic Regulation

A third paradigm of regulation—and the dominant governance mechanism for large social media companies, such as Facebook, Twitter, and YouTube—relies on automated techniques

[19] James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42 (2015); Lampe & Resnick, *supra* note 12.

[20] Michael Bernstein et al., *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*, *in* INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM) 50 (2011).

for evaluating potentially harmful content.[21] This class of strategies uses machine learning and natural language processing to develop computational models that systematically evaluate large quantities of data.

To facilitate scaled content moderation, machine learning models are typically trained to detect language that may be abusive or violent,[22] often automatically removing entities at a certain level of model confidence. Although automated content moderation approaches continue to improve, accurate and reliable detection is challenging at best, even in far less complex applications than the detection of nuanced behaviors like online harassment and hate speech. Social media companies have to make necessary trade-offs between a model's precision (i.e., accuracy) and its recall, or the quantity of relevant instances the model returns. They often optimize for recall out of necessity—nearly a billion tweets are sent per day[23]—resulting in imprecise models plagued by false positives, where harmful content evades detection (where permissible content is incorrectly removed), and true negatives (where harmful content evades detection).

Contrary to popular perception, algorithmic regulation does not eradicate the need for human input. Supervised learning

[21] This has been referred to as the "industrial approach" in Robyn Caplan, *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches,* DATA & SOCIETY (2018).
[22] Eshwar Chandrasekharan et al., *The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data*, *in* PROCEEDINGS OF THE 2017 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 3175 (2017); Hossein Hosseini et al., *Deceiving Google's Perspective API Built for Detecting Toxic Comments* (2017), http://arxiv.org/abs/1702.08138; Ellery Wulczyn, Nithum Thain & Lucas Dixon, *Ex Machina: Personal Attacks Seen at Scale*, in PROCEEDINGS OF THE 26TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 1391 (2017); Dawei Yin et al., *Detection of Harassment on Web 2.0*, *in* PROCEEDINGS OF THE CONTENT ANALYSIS IN THE WEB 2.0 WORKSHOP (2009).
[23] Twitter Usage Statistics, https://www.internetlivestats.com/twitter-statistics/ (last visited May 31, 2021).

models—i.e., a machine learning model that predicts the similarity between a given piece of text and the dataset used to "teach," or train, the model—requires high volumes of annotated data, typically labeled by humans, both for training initial models and for refining their performance over time. Although investing in algorithmic regulation will relieve some burden from workers—companies with well-performing algorithms can, over time, rely on fewer workers for manual content moderation—machine learning still requires a sizeable workforce of human laborers to review hateful, violent, and otherwise traumatizing content over long shifts and for low wages.[24]

Finally, automated governance is also relatively easy to bypass through subtle modifications of language.[25] When combined, these limitations can result in harmful content persisting on social media while jokes, cultural references, and in-group conversations are, from the user's perspective, inexplicably removed.

Retributive Regulation

A fourth paradigm of governance, which has risen to prominence most recently, reflects a complex spectrum of conditions in which social media users aspire to enforce justice themselves—in part due to the recognized failures of social media companies to adequately govern their platforms.[26] When offenders are not held accountable for their actions, users may instead turn to moral shaming to enact retribution[27]—resulting in punishments that, as Kate Klonick argues, may be indeterminate, uncalibrated, or inaccurate.

[24] ROBERTS, *supra* note 3.
[25] Hossein Hosseini et al., *supra* note 22.
[26] Lindsay Blackwell et al., *Classification and Its Consequences for Online Harassment: Design Insights from HeartMob*, 1 PROCS. OF THE ACM ON HUM.-COMPUT. INTERACTION (2017).
[27] JON RONSON, SO YOU'VE BEEN PUBLICLY SHAMED (2016).

An individual user leveraging social media to retaliate against a perceived offender may seem unremarkable; however, the affordances of networked platforms can escalate ordinary social sanctioning into something resembling mass vigilantism. Social feedback (such as likes or upvotes) and algorithmic amplification promote perceptions of endorsement that can result in large-scale group behaviors, which often have extreme and disproportionate impacts on perceived offenders—including threats to physical safety, unwanted disclosures of personal information, sustained social isolation, and job loss.[28]

Retributive regulation is sometimes crudely collapsed into a single set of behaviors, without consideration for the kinds of injustices or harms that necessitate those behaviors. For example, so-called "cancel culture"—a neologism describing a type of mass social sanctioning in which a person's social or professional status is questioned due to a perceived infraction—has arisen as one outcrop of this fourth governance paradigm. Characterizations about the existence of cancel culture should be evaluated cautiously; Meredith Clark argues that the label is often misused, with justifiably critical responses to legitimately harmful acts regularly dismissed as "cancel culture" without recognition of the desired accountability.[29]

This most recent paradigm shift, coupled with the proliferation of online misinformation and increasing political discord, has accelerated demands for formal regulation to hold social media companies accountable for the ramifications of inadequate platform governance. These demands coincide with

[28] RONSON, *supra* note 27; GOLDBERG, *supra* note 7; Citron, *A New Compact for Sexual Privacy*, *supra* note 7.

[29] Meredith D. Clark, *DRAG THEM: A Brief Etymology of So-Called "Cancel Culture"*, 5 COMMC'N & PUB. 88 (2020).

ongoing discussions about the possibilities and limitations for users and communities to regulate themselves.[30]

HARMS DUE TO INADEQUATE SOCIAL MEDIA GOVERNANCE

Harm refers to damage, injury, or setbacks toward a person, entity, or society. Some harms are small and easily repairable, such as the theft of a bicycle. Others, such as the loss of health, are irreparable and cannot be adequately compensated. Harm is distinct from violence, though they are linked; violence will by definition typically cause harm. Harm is a complex and varied concept without a single definition or interpretation; what constitutes harm will vary with use and context. In legal contexts, harm refers to loss or damage to a person's right, property, or well-being, whether physical or mental. In Internet law, scholars have argued for legal recognition of particular kinds of privacy harms,[31] data breach harms,[32] and intimate data harms.[33] Our focus lies in sociotechnical harms—the online content or activity that inflicts psychological or psychological damage towards a person or community and that compromises their ability to participate safely and equitably both online and offline."

Social media platforms facilitate myriad harms, from sexual harassment to hate speech to racism to disinformation. These harms can be intentional (e.g., doxxing a journalist because she wrote something somebody did not like) or unintentional (e.g., sharing content on Twitter that may be inaccessible to disabled people). Intent is a slippery concept to measure; someone intending to be helpful or supportive may still cause harm regardless, in the same

[30] Joseph Seering, *Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation*, 4 PROC. ACM HUM.-COMPUT. INTERACT. 107:1 (2020).

[31] Ryan Calo, *The Boundaries of Privacy Harm Essay*, 86 IND. L.J. 1131 (2011).

[32] Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 TEX. L. REV. 737 (2017).

[33] Citron, *A New Compact for Sexual Privacy*, *supra* note 7.

way that someone who intends to cause harm may claim otherwise when facing undesirable consequences. Additionally, harmful experiences can be differentially traumatic to different people and groups.

We consider two predominant, intersecting categories of harms: platform-perpetrated harms (i.e., those perpetrated by the design of platforms) and platform-enabled harms (i.e., those facilitated by platforms but perpetrated by users or groups). These categories build on our stance that consequences of inadequate platform governance are the responsibility of the platforms themselves.

Psychological Distress

Interpersonal abuse, such as online harassment and hate speech, is widespread and can be profoundly damaging for both targets and bystanders. The effects of harassment vary from person to person, ranging from anxiety, humiliation, and self-blame to anger and physical illness.[34] Online harassment in particular can "cast a long shadow," due in part to the persistence and searchability of digital media—severe harassment can inflict long-term damage to an individual's reputation, comfort, or safety. Perhaps most critically, online harassment has a chilling effect on future disclosures: Lenhart et al. found that, in 2016, 27% of American internet users were self-censoring what they post online due to fear of harassment.[35]

Thus, although harassment is instantiated online, targets of online harassment frequently report disruptions to their offline lives,

[34] Maeve Duggan, *Online Harassment*, PEW RESEARCH CENTER (Oct. 22, 2014), https://www.pewresearch.org/internet/2014/10/22/online-harassment/.

[35] Amanda Lenhart et al., *Online Harassment, Digital Abuse, and Cyberstalking in America*, DATA & SOCIETY (2016), https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/.

including emotional and physical distress, changes to technology use or privacy behaviors, and increased safety and privacy concerns. People who experience harassment often choose to temporarily or permanently abstain from social media sites, despite the resulting isolation from information resources and support networks. Online harassment can also be disruptive to personal responsibilities, work obligations, and sleep due to the labor of reporting harassment to social media platforms or monitoring accounts for activity. Some types of online harassment specifically aim to disrupt a target's offline life, such as swatting (i.e., falsely reporting a crime to encourage law enforcement agencies to investigate a target's home or business).

Online abuse can also result in fear for one's physical safety, regardless of whether or not threats of physical harm ever materialize. Revealing a person's home address, for example, results in a loss of perceived security that endures even after any online harassment has ceased[36]—highlighting the tangible impact of even a potential for harm on the ability for social media users to live safely and comfortably.

Physical Violence

Numerous studies demonstrate the correlation between inciting language online and subsequent offline violence, particularly when social media is used to stoke existing physical conflict. Desmond Patton and coauthors have described the use of social media by gang-involved youth to levy taunts and threats against rival groups, often in response to romantic conflict or expressions of grief and amplified by the affordances of social

[36] *See* stories in GOLDBERG, *supra* note 7.

media platforms.[37] The rapid exchange of comments, pictures, and videos between existing rivals—exacerbated by the network-based visibility of social media content[38]—intensifies any perceived slights, increasing the likelihood of online conflict escalating to physical fights. This perpetuates a cycle of physical and emotional violence in which young people struggling with loss turn to social media for support and instead find themselves embroiled in additional conflict.[39]

Facebook has acknowledged its platform's role in fomenting ethnic violence in Myanmar, in large part due to the deliberate spread of misinformation used to stoke pre-existing tensions between Myanmar's majority-Buddhist population and the Rohingya, a minority Muslim community subjected to ongoing persecution by military and state actors.[40] Despite warnings by researchers and human rights activists about the proliferation of Burmese hate speech on its platform, investigative journalists continued to find hate speech, threats of violence, and calls for genocide on the platform.[41] Similarly, Twitter itself has recognized its role in the January 6, 2021 "storming" of the US Capitol building which resulted in violence, destruction, and fatalities. Soon after the

[37] Desmond Upton Patton et al., *Internet Banging: New Trends in Social Media, Gang Violence, Masculinity and Hip Hop*, 29 COMPUT. IN HUM. BEHAV. A54 (2013); Desmond Upton Patton et al., *You Set Me Up: Gendered Perceptions of Twitter Communication Among Black Chicago Youth*, 6 SOC. MEDIA & SOCIETY (2020); Desmond Upton Patton et al., *Expressions of Loss Predict Aggressive Comments on Twitter Among Gang-Involved Youth in Chicago*, 1 NPJ DIGITAL MEDICINE 1–2 (2018).

[38] Caitlin Elsaesser et al., *Small Becomes Big, Fast: Adolescent Perceptions of How Social Media Features Escalate Online Conflict to Offline Violence*, 122 CHILD. & YOUTH SERVICES REV. 122 (2021).

[39] Patton et al., *Internet Banging*, *supra* note 38.

[40] Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, THE NEW YORK TIMES, (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

[41] Steve Stecklow, *Why Facebook Is Losing The War on Hate Speech in Myanmar*, REUTERS (Aug. 15, 2018), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/; Stevenson, *supra* note 41.

insurrection, and after repeated calls for the removal of inciting tweets by then-President Donald Trump, Twitter permanently removed Trump's account, citing risks of further violence.[42]

Similar violence around the world has been associated with the proliferation of misinformation and hate speech on social media platforms. The circulation of rumors on WhatsApp—an encrypted chat client owned by Facebook—has contributed to a rise in mob lynchings across India.[43] In post-war Sri Lanka, increased violence against Muslim communities and other religious minorities has coincided with an increase in the country's social media users, particularly among Sinhalese Buddhists. [44] In the United States, numerous acts of white supremacist violence were perpetrated by domestic extremists who participated in radical online forums (e.g., Gab, Parler, 4chan).[45] In Pakistan, women are have been silenced through threats of, or actual, violence and death; in 2016, ongoing harassment of Qandeel Baloch, a social media celebrity and activist, culminated in her murder by her own brother.[46]

While threats of physical violence can be delivered on any social media user or community, they often reflect existing disparities between populations: those who are able to exist safely in their homes and local communities may also be able to be safer

[42] *Permanent suspension of @realDonaldTrump*, TWITTER (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html.

[43] Chinmayi Arun, *On WhatsApp, Rumours, Lynchings, and the Indian Government*, 54 ECON. & POL. WKLY. (2019).

[44] Sanjana Hattotuwa, *Digital Blooms: Social Media and Violence in Sri Lanka*, TODA PEACE INSTITUTE, 12 (2018), https://toda.org/assets/files/resources/policy-briefs/t-pb-28_sanjana-hattotuwa_digital-blooms-social-media-and-violence-in-sri-lanka.pdf.

[45] Laurel Wamsley, *On Far-Right Websites, Plans To Storm Capitol Were Made In Plain Sight*, NPR (Jan. 7, 2021), https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/07/954671745/on-far-right-websites-plans-to-storm-capitol-were-made-in-plain-sight.

[46] Imran Gabol & Taser Subhani, *Qandeel Baloch murdered by brother in Multan: police*, DAWN (July 16, 2016), http://www.dawn.com/news/1271213.

online, while those who experience discrimination and persecution offline may be similarly vulnerable online.

Oppression and Marginalization

We cannot talk about harm without also talking about power, because power differences are structural enablers of harm. Power enables abuse through its facilitation of transgressions and its dismantling of accountability. Power differentials manifest in interpersonal contexts (e.g., based on gendered hierarchies)[47] as well as in organizational contexts (e.g., based on workplace hierarchies).[48] Power differentials also arise in emergent ways on social media; influencer status and follower counts provision enormous power to users who gain those statuses or counts,[49] without guidance for or calibration around wielding that power appropriately. Around the world, vulnerable social media users, including dissidents, women, people of color, refugees, transgender people, and members of other non-dominant social groups experience disproportionate harm in online contexts.[50] These experiences are often overlooked, dismissed, or exacerbated by systems of platform governance that fail to account for or even acknowledge the systemic power disparities that enable them.

Technology reflects—and often exacerbates—structural inequities that persist in society writ large. While platforms bear

[47] Christopher Uggen & Amy Blackstone, *Sexual Harassment as a Gendered Expression of Power*, 69 AM. SOCIO. REV. 64 (2004).

[48] *Id.*

[49] TAMA LEAVER, TIM HIGHFIELD & CRYSTAL ABIDIN, INSTAGRAM: VISUAL SOCIAL MEDIA CULTURES (2020).

[50] YORK, *supra* note 2; *Online violence: Just because it's virtual doesn't make it any less real*, GLOBAL FUND FOR WOMEN (2015), https://www.globalfundforwomen.org/online-violence-just-because-its-virtual-doesnt-make-it-any-less-real/; *Toxic Twitter – A Toxic Place for Women*, AMNESTY INT'L (2018), https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.

responsibility for hosting and facilitating harassment, violence, and extremism, these are enduring social problems that cannot be rooted out by social media reform alone. For decades, scholars have documented how racist behavior online intersects with existing offline racism.[51] In 2009, early facial recognition technology developed by HP could easily track the movements of a white user, but failed to recognize black users; later, in 2015, Google's own facial recognition technology categorized pictures of black people as containing images of gorillas.[52] In 2017, despite Apple's efforts to train its own Face ID technology on a large and diverse set of faces,[53] a Chinese woman discovered that her colleague—also a Chinese woman—was able to unlock her device on every attempt.[54] In her book Algorithms of Oppression, Safiya Noble (2018) details countless examples of racial biases that have been "baked in" to the technological systems we use every day: for example, Google returning pictures of white women when queried for images of "professional women," but pictures of black women when queried for images of "unprofessional hair."[55]

Gender and sexual discrimination is also prevalent in technology design, from default avatars registering as male

[51] LISA NAKAMURA, CYBERTYPES: RACE, ETHNICITY, AND IDENTITY ON THE INTERNET (2002); JESSE DANIELS, CYBER RACISM: WHITE SUPREMACY ONLINE AND THE NEW ATTACK ON CIVIL RIGHTS (2009); Gray, *supra* note 18.

[52] Klint Finley, *Can Apple's iPhone X Beat Facial Recognition's Bias Problem?*, WIRED (Sept. 13, 2017), https://www.wired.com/story/can-apples-iphone-x-beat-facial-recognitions-bias-problem/.

[53] Kate Conger, *How Apple Says It Prevented Face ID From Being Racist*, GIZMODO (Oct. 16, 2017), https://gizmodo.com/how-apple-says-it-prevented-face-id-from-being-racist-1819557448.

[54] Christina Zhao, *Is the iPhone X's facial recognition racist?*, NEWSWEEK (Dec. 18, 2017), https://www.newsweek.com/iphone-x-racist-apple-refunds-device-cant-tell-chinese-people-apart-woman-751263.

[55] SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

silhouettes[56] to Facebook's ongoing challenges surrounding its "real name" policy and the deactivation of accounts belonging to trans users, drag queens, Indigenous people, abuse survivors, and others whose identities or account names may be inconsistent with their legal names.[57] Most online forms requiring gender information only offer a binary choice—"male" or "female"—forcing non-binary individuals to either choose an incorrect gender category or refrain from using the site or service.[58] The implicit biases designed into everyday technologies not only reflect existing discrimination, but may also exacerbate it: exposure to negative stereotypes about one's social identity can actually reduce performance on a relevant task, a phenomenon known as stereotype threat.[59] Further, these technological biases, however unintentional, are often only identified—and subsequently given the opportunity for correction—through proactive auditing by researchers, in a practice Sandvig, et al. (2014) call algorithmic auditing.[60]

These challenges are partly, though not entirely, due to problems of classification. Social media platforms rely on numerous

[56] April H. Bailey & Marianne LaFrance, *Anonymously Male: Social Media Avatar Icons Are Implicitly Male and Resistant to Change*, 10 CYBERPSYCHOLOGY: J. PSYCH. RSCH. ON CYBERSPACE (2016).

[57] Vauhini Vara, *Drag Queens Versus Facebook's Real-Names Policy*, THE NEW YORKER (Oct. 2, 2014), https://www.newyorker.com/business/currency/whos-real-enough-facebook; Oliver L. Haimson & Anna Lauren Hoffmann, *Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities*, 21 FIRST MONDAY (2016).

[58] Scheuerman, Morgan Klaus et al., *Revisiting Gendered Web Forms: An Evaluation of Gender Inputs with (Non-) Binary People*, *in* PROCEEDINGS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (May 2021).

[59] Claude M. Steele, Steven J. Spencer & Joshua Aronson, *Contending with Group Image: The Psychology of Stereotype and Social Identity Threat*, 14 ADVANCES IN EXPERIMENTAL AND SOC. PSYCH. 379 (2002).

[60] Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, in* DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY (2014).

classification systems and categorization schema[61]: algorithmic feeds serve specific content based on particular features; reporting flows ask users to identify specific policy violations; profile creation requires various selections from predefined lists. But when classification systems are built to optimize for scale, variation is flattened in favor of majority experiences. This results in compounding harms for users and communities who are already socially, economically, or otherwise excluded from society. For example, when sex trafficking is prohibited on mainstream platforms, consensual sex work is often caught up in the same algorithmic net; this has the immediate material effect of reduced income for sex workers (who themselves often possess multiple stigmatized identities such as being queer or non-white), while also contributing to the continued stigmatization of sex-based labor.[62] The embedded biases inherent in large-scale automation manifest in many forms, across gender, race, disability, and other characteristics—most acutely at their intersections—and often in ways that are not transparent or interpretable to the users whose experiences are governed by them.

Threats to Free Expression

Regulatory recommendations typically focus on refinements to specific legislation. In the U.S., scholars have called for "reasonable moderation practices rather than the free pass" that is enabled by 47 U.S.C. § 230, a provision of the Communications Decency Act (CDA) of 1996 protecting online service providers from incurring legal liability for third-party (i.e., user-generated)

[61] Blackwell et al., *Classification and its Consequences for Online Harassment*, *supra* note 13.

[62] See stories from sex workers documented in Kendra Albert et al., *FOSTA in Legal Context* (2020), https://papers.ssrn.com/abstract=3663898; YORK, *supra* note 2.

content.[63] Platforms frequently cite freedom of expression when deciding to minimize their role in arbitration, a stance buttressed by the "safe harbor" offered by Section 230.[64]

Though Section 230 has had an outsized influence on US-based corporate governance, many regions around the world are debating regulatory practices, with varying thresholds for the types of content social media companies are legally required to remove. In Germany, NetzDG requires platforms to promptly remove illegal content in Germany, including Anti-Semitic speech and hate speech based on religion or ethnicity.[65] In Korea, Article 44 of the Information and Communications Network Act (ICNA) encourages proactive removal of content if requested.[66] In India, the IT Act provides immunity for platforms as long as they take action to address certain categories of content within a short time frame.[67] In Australia, platforms have to moderate and also report "abhorrent violent" content.[68] In other countries, such as Syria, Turkey, Pakistan, and Tunisia, partial or wholesale bans on social media result in widespread censorship of expression by state actors.[69]

[63] DANIELLE KEATS CITRON & MARY ANNE FRANKS, *The Internet As a Speech Machine and Other Myths Confounding Section 230 Reform*, U. CHI. L. FORUM (forthcoming, 2020).
[64] *Id*.
[65] *Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act, NetzDG) - Basic Information*, BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ [FEDERAL MINISTRY OF JUSTICE AND CONSUMER PROTECTION], https://www.BMJV.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html (last visited Apr 8, 2021).
[66] *Act on Promotion of Information and Communications Network Utilization and Information Protection, etc.*, KOREAN LAW TRANSLATION CENTER, https://elaw.klri.re.kr/eng_service/lawView.do?hseq=38422&lang=ENG (last visited Apr 8, 2021).
[67] The Information Technology Act, 2000 (India).
[68] Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act, 2019 (Austl.).
[69] For a comprehensive discussion of platforms, free speech and censorship, and state governance, see YORK, *supra* note 2.

Freedom of expression is a human right; however, its contours are nuanced and vary by regions and contexts (e.g., attitudes towards nudity, which is considered normative in some cultures but highly sensitive in others). Preserving freedom of expression while also mitigating harm is a complex endeavor. For example, in her book, *Silicon Values*, Jillian York highlights how platforms' automated removal of violent extremist content prompted human rights groups to begin preserving that content as evidence of war crimes.[70] Chinmayi Arun notes that mounting pressure on social media companies to cooperate with governments has alarming implications—both for individual user privacy and the continued utility of these platforms for journalists, activists, and political dissidents.[71] While this article is not focused on the specific nuances of free expression, any proposal for shifts in social media governance must also consider implications for human rights, including the potential for exploitation by state actors.

PRINCIPLES FOR SOCIAL MEDIA GOVERNANCE

Although social media governance to date has largely been informed by Western models of criminal justice, which rely on sanctions (e.g., punishment) to encourage compliance with formal rules and laws, we argue for systems of governance that instead focus on accountability for and repair of specific harms. Social media governance should be informed by both punitive and restorative frameworks; here, we propose how theories of justice can inform social media policies, practices, and products that acknowledge and attend to harm.

[70] *Id.*

[71] Chinmayi Arun, *Facebook's Faces*, 135 HARV. L. REV. F. (forthcoming).

Retribution and Punishment

The concept of justice is invoked when deciding how society should respond to a person who is perceived to have committed some infraction (i.e., a violation of rules and laws). In Western societies, criminal justice approaches have traditionally sought to discourage offenders through the fear of strict criminal sanctions. The concept of retribution is focused on delivering offenders their deservedness,[72], and proportionality[73] in criminal sentencing. Moral judgment plays a powerful role in retribution and shapes cultural attitudes, policy, and law around appropriate punishments.[74] Feelings of moral anger and disgust (e.g., feelings that results if someone engages in pedophilia) often protect and preserve social order within a society.[75] In the United States, incarceration has been a predominant engine for enacting punishment, especially towards some groups including people of color, disabled people, and poor people.[76]

Social media governance has typically adopted Western frameworks of criminal justice: identifying perpetrators of undesirable behavior and administering punitive responses.[77] If

[72] Kevin M. Carlsmith & John M. Darley, *Psychological Aspects of Retributive Justice*, 40 ADVANCES IN EXPERIMENTAL SOC. PSYCH. 193 (2008); IMMANUEL KANT & WERNER PLUHAR, CRITIQUE OF JUDGMENT (1987).

[73] Michael Wenzel et al., *Retributive and Restorative Justice*, 32 L. HUM. BEHAV. 375 (2008).

[74] Roger Giner-Sorolla et al., *Emotions in Sexual Morality: Testing the Separate Elicitors of Anger and Disgust* 26 COGNITION & EMOTION 1208 (2012); Jesse Prinz, *Is Morality Innate?, in* MORAL PSYCHOLOGY: THE EVOLUTION OF MORALITY: ADAPTATIONS AND INNATENESS 608 (Walter Sinnott-Armstrong & Christian B. Miller eds., 2007).

[75] Bunmi O. Olatunji & Craig N. Sawchuk, *Disgust: Characteristic Features, Social Manifestations, and Clinical Implications*, 24 J. SOC. & CLINICAL PSYCH. 932 (2005); Pascale Sophie Russell & Roger Giner-Sorolla, *Moral Anger, but Not Moral Disgust, Responds to Intentionality*, 11 EMOTION 233 (2011).

[76] RUEBEN JONATHAN MILLER, HALFWAY HOME (2021).

[77] Bradford et al., *supra* note 2; Eshwar Chandrasekharan et al., *supra* note 2; Shagun Jhaver, Amy Bruckman & Eric Gilbert, *Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations*

content is found to violate a platform's community guidelines, platform responses range from removing the content or demoting its visibility to banning the user who produced it, either temporarily or permanently. However, these sanctions embrace many of the problems of retributive models of governance; namely, they overlook the needs and interests of the targets of harassment and remove offenses and offenders from the community without any attempt at rehabilitation. Contemporary platform governance also relies on obfuscated processes of content moderation that have little transparency or accountability to all involved parties[78]; content is deleted without leaving any visible trace of its removal; policy violators have little opportunity for recourse and may not even be informed of the specific rule they have broken; reporters receive generalized responses that often don't reference the content in question, if they receive a response at all.

In typical platform-driven moderation systems, all violators are treated equally, with users who unintentionally violate rules receiving the same sanctions as users who deliberately try to cause harm. Instead, we argue for an expanded set of remedies, one that better recognizes and remediates harms by incorporating responsive penalties that allow for reeducation, rehabilitation, and forgiveness. Social media users already intuitively imagine diverse and varying punishments that allow for proportional responses to varied infractions, depending both on the specific type of violation and the perceived intent of the violator.[79] For example, people who

on Reddit, CSCW PROC. ACM HUM.-COMPUT. INTERACT. (2019); J. Nathan Matias, *supra* note 15; Pater et al., *supra* note 13; Sarah Perez, *Twitter adds more anti-abuse measures focused on banning accounts, silencing bullying*, TECHCRUNCH (Mar. 1, 2017), http://social.techcrunch.com/2017/03/01/twitter-adds-more-anti-abuse-measures-focused-on-banning-accounts-silencing-bullying/.

[78] The Santa Clara Principles, *supra* note 2.

[79] Lindsay Blackwell et al., *Harassment in Social Virtual Reality: Challenges for Platform Governance*, 3 PROC. ACM HUM.-COMPUT. INTERACT. 100:1 (2019).

perpetuate one-time or occasional offenses can be given the opportunity to correct and make amends for their behavior, with more severe penalties reserved for users who perpetuate sustained abuse without remorse.

Moderation practices that eschew blunt, one-size-fits-all penalties in favor of sanctions which are proportionate to specific violations is aligned with what Braithwaite calls responsive regulation.[80] In a responsive regulation framework, the least interventionist punishments—for example, education around existing rules and policies—are applied to first-time or other potentially redeemable offenders, with sanctions for repeat violators escalating in severity until they reach total incapacitation (e.g., a permanent account- or IP address-level ban).[81] By implementing enforcement decisions that are responsive to the context of specific infractions, platforms may be perceived as more legitimate when harsher penalties are required: a user won't become eligible for permanent suspension without being given multiple opportunities to correct their behavior and adhere to platform policies. Responsive regulation may also help platforms avoid alienating users for incorrect enforcement decisions; when the full context surrounding a violation is unclear, a less severe penalty can be applied.

Accountability and Restoration

Alternative justice models for platform governance could recognize harm, establish accountability for that harm, and establish an obligation to repair harm. Whereas a retributive justice governance approach would ask what laws have been broken, who broke them, and what punishment is deserved, alternative justice

[80] IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION: TRANSCENDING THE DEREGULATION DEBATE (1992).

[81] JOHN BRAITHWAITE, RESTORATIVE JUSTICE & RESPONSIVE REGULATION (2002).

approaches would instead ask who has been harmed, what do they need, and how should systems be redesigned to prevent harms from reoccurring? However, alternative justice systems are not in themselves sufficient to address harm; any justice system that is implemented—whether traditional or alternative—may inadvertently protect and benefit social groups who are already privileged unless the systems are explicitly designed to do otherwise.

Two prominent alternative justice frameworks are restorative justice and transformative justice. Restorative justice is a framework and movement that encourages mediated conversations between those who perpetuate and those who experience harm, typically with mediators and community members actively participating. Restorative justice asks that offenders acknowledge wrongdoing and harm, accept responsibility for their actions, and express remorse. Restorative justice has been practiced in Indigenous communities, and has been advanced as an alternative to Western criminal justice systems that over-incarcerated Indigenous youth. In New Zealand, for example, restorative justice was the foundation for a 1989 act between Maori people and New Zealand Parliament which was designed to care for Indigenous children rather than moving them into prison pipelines.[82]

Recognition of wrongdoing is an essential first step in establishing accountability for harm. The concept of recognition is often invoked in human rights discussions and contains two facets: recognition of human rights, and recognition of violations of those rights. However, recognition has also been misused as a politicized form of collective identity that demands recognition of a dominant group while perpetuating distributive injustices towards non-

[82] The Oranga Tamariki Act, 1989 (N.Z.).

dominant groups.[83] Restorative justice programs were sometimes implemented without consideration of race or disability;[84]; as a result, able bodied white women offenders might have been viewed as victims of circumstance who deserved empathy, while disabled people of color continued to be over-incarcerated.[85] Many restorative justice practitioners have chosen to work outside of criminal legal systems because of the ongoing failures of those systems. Thus, recognition is not simply a decision to acknowledge harms, but a confluence of decisions about what rights people should have, how to acknowledge those rights, and how to acknowledge violations of those rights.

Recognition of harm on social media asks for recognition of the multitudes of ways that users and communities can experience harms, including those that fall outside of current regulatory capture. Accountability, then, requires accepting responsibility for those harms, including the obligation to repair them. Scholars Mia Mingus and Mariame Kaba have argued for moving away from holding others accountable and towards supporting proactive accountability, i.e., "active accountability."[86] Centering accountability and repair

[83] Nancy Fraser, *Rethinking Recognition: Overcoming Displacement and Reification in Culture Politics*, *in* RECOGNITION STRUGGLES AND SOCIAL MOVEMENTS: CONTESTED IDENTITIES, AGENCY AND POWER (2003).

[84] Theo Gavrielides, *Bringing Race Relations Into the Restorative Justice Debate: An Alternative and Personalized Vision of "the Other"*, 45 J. BLACK STUD. 216 (2014).

[85] Danielle Dirks et al., *'She's White and She's Hot, So She Can't Be Guilty': Female Criminality, Penal Spectatorship, and White Protectionism*, 18 CONTEMP. JUST. REV. 160 (2015).

[86] Mariame Kaba et al., *When It Comes to Abolition, Accountability Is a Gift*, BITCH MEDIA , https://www.bitchmedia.org/article/mariame-kaba-josie-duffy-rice-rethinking-accountability-abolition (last visited Jan 6, 2021); Mariame Kaba & John Duda, *Towards the Horizon of Abolition: A Conversation With Mariame Kaba* (2018), https://transformharm.org/towards-the-horizon-of-abolition-a-conversation-with-mariame-kaba/ (last visited Jan 8, 2021); Mia Mingus, *The Four Parts of Accountability: How To Give A Genuine Apology Part 1*, LEAVING EVIDENCE (Dec. 18, 2019), https://leavingevidence.wordpress.com/2019/12/18/how-to-give-a-good-apology-part-1-the-four-parts-of-accountability/.

requires shifts towards the needs of those harmed, and accountability from those who perpetuate harm. Acts like apologies, mediated conversation, proclamations, and commemorations could all be supported in online interactions as non-material forms of restoration and accountability.[87] For example, apologies can be powerful illocutionary devices for amending wrongdoings, though they need to be genuine or they can further magnify harm, especially for groups who have already experienced oppression.[88] Similarly, intent not to commit harm again, and subsequent actions, can be a form of accountability and restoration. These boundaries could be built into the design of social media sites where targets of harassment could be granted agency to decide whether to engage further, and if so, under what terms. Other acts like compensation or amplification could enact material remedies, which may be important for correcting some kinds of online injustices. While accountability processes hopefully result in resolution, that may not always be attainable, and the burden of reaching resolution should not fall on those who have experienced harm.[89]

Transformative justice, which extends restorative principles and practices beyond individual reconciliation and towards

[87] Our prior studies show that U.S. adults and young adults are generally favorable towards the idea of apologies after online harassment. *See* Sarita Schoenebeck, et al., *Drawing from Justice Theories to Support Targets of Online Harassment*, 23 NEW MEDIA & SOCIETY 1278 (2020); Sarita Schoenebeck et al., *Youth Trust in Social Media Companies' Responses to Online Harassment*, PACM HUM.-COMPUT. INTERACTION 2:1 (2021).

[88] Schoenebeck et al., Drawing from Justice Theories to Support Targets of Online Harassment, supra note 88; Schoenebeck et al., *Youth Trust in Social Media Companies' Responses to Online Harassment, supra* note 88. While apologies can be a conduit for justice, the delivery of an apology should not create an expectation of forgiveness from the target, nor should it imply that accountability was present.

[89] John Braithwaite, *Restorative Justice: Assessing Optimistic and Pessimistic Accounts*, 25 CRIME & JUSTICE 1 (1999); Jung Jin Choi, Gordon Bazemore & Michael J. Gilbert, *Review of Research on Victims' Experiences in Restorative Justice: Implications for Youth Justice*, 34 CHILD. & YOUTH SERVS. REV. 35 (2012).

systematic change, has been similarly developed and advanced by non-dominant social groups, including immigrant, Indigenous, Black, disabled, and queer and trans communities.[90] Transformative justice involves practices and politics focused on ending sexual violence using processes outside of carceral policing systems. Transformative justice movements propose that prison and state systems create more harm, violence, and abuse rather than addressing them. Two tenets are that violence and abuse should be responded to within communities rather than by criminal legal systems (while noting that communities themselves can also perpetuate violence), and that any responses should combat, rather than reinforce, oppressive societal norms. Transformative justice movements seek not only to respond to current violence, but to address cycles of violence by transforming the conditions that allowed it to happen.

While restorative justice and transformative justice are distinct movements with different principles, they share a commitment to recognizing harm and violence and resisting the carceral systems that perpetuate them. These commitments help to shed light on the failures of current platform governance practices; when platforms fail to explicitly acknowledge and combat existing inequity, they further entrench those harms with content moderation policies that may seem appropriate on an individual level (e.g., disallowing hate speech), but which obscure and perpetuate violence at a structural level (e.g., equating hate speech against men with that against women, which overlooks gender-based oppression). Many

[90] BEYOND SURVIVAL: STRATEGIES AND STORIES FROM THE TRANSFORMATIVE JUSTICE MOVEMENT (2020); Sara Kershnar et al., *Toward Transformative Justice*, GENERATION FIVE (2007), http://www.usprisonculture.com/blog/wp-content/uploads/2012/03/G5_Toward_Transformative_Justice.pdf; Mia Mingus, *Transformative Justice: A Brief Description*, LEAVING EVIDENCE (Jan. 9, 2019), https://leavingevidence.wordpress.com/2019/01/09/transformative-justice-a-brief-description/.

approaches to platform governance can be characterized as "reformist reforms"[91] a term for reforms which maintain the status quo by upholding existing oppression systems. In policing, non-reformist reforms would include those that reduce, rather than maintain, the power by police themselves; reformist reforms would be those which instead increase police funding (e.g., body cameras) or scale (e.g., community policing), effectively maintaining or even strengthening the existing systems. Content moderation discussions can easily fall into reformist reform traps—they tweak, tune, and slightly improve what content is moderated and how, while cementing in place governance structures that continue to overlook harms.

Repairing harms is not one-size-fits-all, however; different harms may be paired with different frameworks and approaches, and multiple approaches could be combined together.[92] Any design-centered approach must be recognizant of its own limitations; much as a school cannot overcome economic inequality or a prison cannot overcome racism, design cannot repair the underlying systemic injustices it facilitates. Instead, like restorative and transformative justice movements in schools and prisons, design as a praxis should aim to acknowledge and mitigate harms within those sites, while also questioning the underlying systems that enable those harms. Any system of justice—whether traditional or alternative—may inadvertently protect and benefit social groups who are already privileged unless they are explicitly designed to do otherwise.

[91] Kaba & Duda, *supra* note 87.

[92] Eric Goldman, *Content Moderation Remedies*, MICHIGAN TECH. L. REV. (forthcoming), https://papers.ssrn.com/abstract=3810580 (last visited Mar 31, 2021).

Principles for Repairing Harms

We propose several key shifts for social media companies to facilitate the design and development of platform governance models centered on the recognition and repair of harm.

From Neutral to Principled

Social media companies have typically adopted a "neutral" stance, embracing a veneer of impartiality that ostensibly serves to absolve them of the responsibility to adjudicate harm. This aspirational objectivity may be buttressed by an orientation toward measurement, labels, classification, and formalization in how technology is produced.[93] Yet platforms already arbitrate countless decisions, simply by having and enforcing policies for acceptable behavior.[94] Companies make principled decisions about what is included or omitted in their policies or procedures, and they enact those principles whenever they enforce (or choose not to enforce) them. Instead of clinging to the myth of neutral arbitration, platforms should recognize the power they wield—and the values and principles already evident in the decisions they make every day—and move toward explicitly principled governance.

The philosopher and critical theorist Nancy Fraser proposes that accountability for harms involves "seek[ing] institutional remedies for institutionalized harms."[95] Principled governance requires transparency, accountability, and opportunities for appeal[96]—values which are central in theories of procedural justice,

[93] See the field of science and technology studies for a review of how science is produced, e.g., BRUNO LATOUR & STEVE WOOLGAR, LABORATORY LIFE : THE SOCIAL CONSTRUCTION OF SCIENTIFIC FACTS (1979); GEOFFREY C. BOWKER & LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES (1999).
[94] GILLESPIE, *supra* note 2.
[95] Fraser, *supra* note 84.
[96] The Santa Clara Principles, supra note 2.

or the notion that fair and transparent decision-making processes result in more equitable outcomes and, in turn, more cooperative behavior.[97] Some social media companies have begun to respond to public concern about procedural fairness, implementing systems for appealing content removal decisions and experimental initiatives like Facebook's controversial Oversight Board, a group of experts with the authority to overturn a selection of appealed content moderation decisions.[98]

However, principled governance also requires interrogating the limitations of concepts like fairness, despite—or because of—their deep embeddings in many justice systems. Power differences explain why concepts like fairness can overlook injustices: fairness maintains power differentials because it locates the source of problems within individuals or technologies instead of as systemic and contextual inequities.[99] As such, we propose that social media governance must be principled rather than neutral, and that a principled approach requires platforms to reckon with their role in enabling, or magnifying, structural injustices.

From Equality to Equity

Social media companies have traditionally built their policies and procedures around equality, or the notion that all people deserve equal treatment. But equal treatment—which many people may, on its face, consider to be fair—is typically engaged on an individual level, rather than contextualized in a larger system of sociohistorical relationships and systemic injustice. In other words,

[97] BRADFORD ET AL., *supra* note 2. *See* Badeie et. al, this issue.

[98] For an in-depth analysis see Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L. J. 2418 (2019).

[99] Anna Lauren Hoffmann, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFO., COMMC'N & SOCIETY 900 (2019).

while equality aims to promote justice and fairness, it can only work if everybody starts with the same resources and needs. In practice, an equality-based approach—when applied to inequitable systems—only serves to uphold existing systems of oppression and perpetuate systemic inequality, such as racism and transphobia. Most (if not all) social media companies apply their policies using policies of equality, thereby perpetuating equalities rather than remediating them.

For example, Facebook's Community Standards define hate speech as "a direct attack"[100]—described as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation—"against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease." The policy is delineated by different types of attacks, but it applies equally to all groups: a dehumanizing statement against men (e.g., "Men are trash") is treated the same as a dehumanizing statement against women (e.g., "Women are trash"), despite structural sexism (i.e., systematic gender inequality, one manifestation of which is the wage gap[101]).

Thus, while "equal treatment" may seem appropriate on an individual level, it obscures—and ultimately perpetuates—existing inequalities at the structural level. Women, queer people, people of color, dissidents, religious minorities, lower caste groups, and other

[100] *Facebook Community Standards on Hate Speech*, FACEBOOK, https://www.facebook.com/communitystandards/hate_speech (last visited Apr. 4, 2021).

[101] Nikki Graf, Anna Brown & Eileen Patten, *The Narrowing, but Persistent, Gender Gap in Pay*, PEW RESEARCH CENTER (Mar. 22, 2019), https://www.pewresearch.org/fact-tank/2019/03/22/gender-pay-gap-facts/.

groups are disproportionately affected by online harassment [102], particularly when those identities intersect (e.g., a Black trans woman). Why would we expect social media companies to police harassment of these groups with the same fervor—or to detect it at the same volume—as the less frequent and typically lower-severity harassment of their socially-dominant counterparts? Instead, we argue that social media governance should prioritize equity, or the fair distribution of benefits, resources, or outcomes. This is best understood as a question of distributive justice: whereas equality mandates that everyone is given the same resources or opportunities, an equitable approach recognizes that individual circumstances may require uneven distribution in order to ultimately reach an equal outcome.

Because social differences between people (e.g., race) shape what kinds of harm they might experience (e.g., racism), appropriate responses to harm should be interpreted in the broader cultural and social contexts in which the harm occurred. Although behaviors like online harassment manifest as interpersonal conflict, social media platforms contribute to and perpetuate inequities that result in disproportionate harm to vulnerable populations. To successfully recognize and repair harm, social media companies must first address their role in enabling and exacerbating existing structural injustice.

[102] Shawna Chen, Bethany Allen-Ebrahimian, *Harassment of Chinese dissidents was warning signal on disinformation*, AXIOS (Jan. 12, 2021), https://www.axios.com/chinese-dissidents-disinformation-protests-7dbc28d7-68d0-4a09-ac4c-f6a11a504f7c.html; Maeve Duggan, *1 in 4 black Americans have faced online harassment because of their race, ethnicity*, PEW RESEARCH CENTER (Jul. 25, 2017), https://www.pewresearch.org/fact-tank/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/; Duggan, *supra* note 35; Emily Vogels, *The State of Online Harassment*, PEW RESEARCH CENTER (Jan. 13, 2021), https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

From Content to Behavior

Social media companies currently evaluate potentially harmful behavior purely at the content level—that is, content moderators are asked to consider the specific words used in a given post or comment, divorced from contextual factors such as who the author is; who the audience or target is; what the relationship between the author and their audience is, and so on.

While human content moderators will intuit some amount of context from the content itself—for example, a tweet that contains profanity but also a playful emoji may be interpreted as banter between friends—algorithmic (i.e., computational) moderation still cannot. Scaled moderation relies almost exclusively on natural language processing and other machine learning techniques; a typical supervised learning model will be trained on a broad corpus of content and produce blunt, binary judgments—e.g., violating or not violating; hate speech or not hate speech—based on how closely an object resembles the training data set. This results in enforcement outcomes which are almost entirely based on isolated pieces of content, devoid of the sociohistorical context in which they were produced.

Complex and inherently social behaviors like online harassment cannot be understood separate from the context in which they occurred. While the core experience of online harassment may be largely universal across regions and cultures, how people experience harm may vary by individual, context, and culture. For example, non-consensual sharing of intimate images is an intense invasion of privacy regardless of the target’s location—but for women in Pakistan or Saudi Arabia, an intimate image could bring shame to an entire family, creating additional consequences and intensifying an already acutely harmful experience.

A focus on behavior allows for more nuance in what sanctions are applied to potential violators. While dominant models of social media governance typically favor blunt punishments that escalate in severity (e.g., limiting a violator's account privileges for one day after their first violation, three days after a second infraction, and so on), this approach has several limitations. First, applying the same punishment to all policy violators, regardless of the infraction, collapses a wide range of behaviors into a binary determination of "violation" or "no violation." In addition to creating uncomfortably disproportionate outcomes—someone who reacts with justifiable hostility to an instance of racism, for example, will endure the same punishment as someone who posts something racist—this approach does not allow for accountability that more appropriately addresses the root cause of specific behaviors.

Content-centric approaches to social media governance also do not account for differences in what motivates individuals to participate in abusive behavior. While the resulting harm is ultimately the same regardless of the perpetrator's intent, considering the underlying motivation for a behavior allows for more strategic and targeted interventions that may reduce the likelihood of reoffense. For example, a user who is new to a specific social media site may benefit from educational interventions that help the user acclimate to platform rules and norms; a user who engages in retributive harassment is likely aware that they are violating a rule. That user could be prompted to report the person they are seeking to sanction instead.

Finally, content removal is an inherently reactive governance strategy; by the time a post is reported to or reviewed by the platform, it has likely already caused significant harm. Reactive governance is a losing game: users are producing content much faster than platforms can moderate it, no matter how many algorithms they build or moderators they hire. Shifting focus from removing individual pieces of content toward understanding and addressing the underlying behaviors will allow social media platforms to become more proactive in their governance, implementing interventions that discourage harmful behaviors before they manifest on the platform.

From Retribution to Rehabilitation

While criminal justice is an accessible metaphor, it is not a desirable approach to social media governance for a variety of reasons—not least because it privileges a carceral approach that focuses on punishing, rather than rehabilitating, offenders. Retributive governance seeks to restore justice by giving the offender their "just deserts," or a punishment proportional to the offense. While this approach accounts for the severity of harm inflicted, it does nothing to redress the harm itself—in other words, it focuses on the perpetrators of harm, with little to no consideration for the experiences of those who were harmed.

In order to appropriately repair harm, we must first transform social media governance from a system of retribution toward one of accountability. We can draw inspiration from principles of restorative justice, which first asks the injured party to identify their desired path forward. Often, this includes asking the offender to take active accountability for the harm they have caused. Rather than incarcerating offenders, a restorative justice approach seeks to rehabilitate offenders and reintegrate them into the community, reducing the likelihood of recidivism.

This is not to say that punishment is never appropriate. A focus on rehabilitation over punishment allows platforms to better distinguish users who intend to cause harm from those who don't—a distinction many community members already make, particularly in smaller online communities where moderators frequently interact directly with other users.[103] While good intentions may not lessen any resulting harm, they help indicate an appropriate response. On social media, as in offline contexts, a small number of frequent offenders produce a disproportionate amount of violations; some motivated by extrinsic factors (e.g., financial gain) and others by behaviors associated with violence and manipulation.[104] When offending users are given opportunity to correct and make amends for their behavior, more severe penalties, such as IP address-based or device-level bans, can eventually be applied with more legitimacy. This allows platforms to lessen the intensity of negative experiences caused by incorrect enforcement decisions (e.g., model false positives) while also ensuring that extreme offenders are met with swift punitive responses—resulting in safer, more equitable online spaces.

From Authority to Community

We encourage social media platforms to transition away from paternalistic, top-down models of governance in favor of giving communities more control over their own experiences. One reason for this approach is practicality: these are extremely difficult problems that will take years, if not decades, to solve. Online audiences are disparate and often invisible—even to platforms themselves—making it difficult to reliably assess the targets, scope,

[103] Blackwell et al., *Harassment in Social Virtual Reality: Challenges for Platform Governance*, *supra* note 80.

[104] Extensive studies by Neumann and colleagues suggest that about 1% of the population exhibits what has been called psychopathy; however, the psychopathy diagnosis has been contested as overlooking a range of experiences (e.g., disabilities that may falsely present as psychopathy) and should be considered cautiously. Craig S. Neumann & Robert D. Hare, *Psychopathic Traits in a Large Community Sample: Links to Violence, Alcohol Use, and Intelligence*, 76 J. CONSULTING & CLINICAL PSYCH. 893 (2008); Craig S. Neumann et al., *Psychopathic Traits in Females and Males Across the Globe*, 30 BEHAV. SCIS. & L. 557 (2012).

and severity of harms. Platforms are often responsible for evaluating interactions without the necessary context; even when context is available, it is incredibly hard, if not impossible, to evaluate consistently at the scale required to train a machine learning model. We also can’t rely on human moderation alone; while automated enforcement has significant limitations, content moderation is incredibly taxing on workers, who spend every day reviewing the worst of humanity for extremely low wages.

Beyond the practicality of more bottom-up, community-driven governance, giving communities increased agency ultimately reduces harm, both by empowering people to exert control over their own experiences and by creating opportunities for more nuanced, individualized interventions. Increased user agency also helps mitigate the challenges of platforms’ traditional, “one-size-fits-all” approach to global governance: when communities experiencing harm have control over their experiences on the platform, they can decide what justice looks like on their own terms.

Finally, the transition from authority to agency is necessary for decentralizing the incredible amount of power social media companies now wield. Current approaches to social media governance are fundamentally authoritarian; companies exert total control over their content moderation processes, with little to no transparency into how policies are developed, how moderators make decisions, how algorithms are trained, and every other facet of this incredibly complex ecosystem. Social media platforms exist to serve social functions: relationship-building, free expression, collective organizing. We deserve radical transparency into how this small handful of American companies is choosing to govern what are now our primary social spaces.

CONCLUSION

Despite early optimism about social media's democratic promises, social media platforms have enabled abuse and amplified existing systemic injustices. Models of governance that may have sufficed in early, online communities are ineffective at the scale of many contemporary platforms, which largely rely on obscure but powerful automated technologies. Failures to effectively govern platforms manifest in severe consequences for social media users, including psychological distress, physical violence, and the continued suppression of non-dominant voices. Unfortunately, platforms' reproduction of punitive models of governance focus on removing offenders rather than repairing harm. We argue that platforms are obligated to repair these harms, and that doing so requires reimagining governance frameworks that accommodate a wider range of harms and remedies. We propose a set of governing principles to better equip social media companies for accountability to their users.

INFORMATIONAL QUALITY LABELING ON SOCIAL MEDIA: IN DEFENSE OF A SOCIAL EPISTEMOLOGY STRATEGY

John P. Wihbey, Matthew Kopec & Ronald Sandler[*]

INTRODUCTION

Labeling is a content moderation tool that social media companies have at their disposal to indicate to users something about the quality of information that appears on their platforms. Information quality labeling can be either negative or positive. Negative labeling indicates to users that the information they are viewing is of poor or questionable quality—e.g., unverified, false, contested, from an untrusted source. Positive labeling indicates to users that the information they are viewing meets a standard of quality—e.g. verified, fact checked, from a trusted source. Social media companies often deploy negative labeling tactically. That is, moderators use the tool in order to address a particular type of problem as it arises.

For example, prior to the 2020 presidential election Donald Trump indicated that he was likely to declare victory prematurely.

[*] John P. Wihbey, Professor of Journalism and New Media, Northeastern University; Matthew Kopec, Associate Director of the Ethics Institute, Northeastern University; Ronald Sandler, Professor of Philosophy, Chair of Department of Philosophy and Religion, Director of Ethics Institute, Northeastern University. The authors thank Gabriela Compagni, Kay Mathiesen, Nicholas Miklaucic, Garrett Morrow, Roberto Patterson, Jessica Polny, and Briony Swire-Thomson for their contributions to the content labeling project from which this paper was developed and Don Fallis both for his contributions to that project and for helpful comments on an earlier draft. They also thank Sudhir Venkatesh, Seth Cole, Farzaneh Badiei, the editors and reviewers of this journal, and the audience at "News and Information Disorder in the 2020 Presidential Election," sponsored by The Information Society Project at Yale Law School, held on Dec. 4, 2020, for their helpful comments, questions, and suggestions. The work in this paper was supported in part by funding from the Northeastern University Provost's Office and a gift from Facebook.

Late on election night he did just that, and falsely claimed that the election was being "stolen," when in fact legitimate votes were still being counted. Twitter, Facebook, and YouTube labeled Trump's false claims, which after the election continued on topics such as alleged voter fraud in various U.S. states. This moderation pattern continued until the platforms ultimately froze or removed his accounts in the wake of the U.S. Capitol attacks that his social media activities—false claims about the election, promulgation of conspiracy theories, approval of white nationalist extremists, and exhortations to fight the outcome—helped to foment.[1] The platforms also have used information quality labeling as part of the effort to prevent the spread of COVID-19 misinformation, QAnon conspiracy theories, and mail-in voting misinformation, for example.[2] The use of labeling in these contexts is tactical in the sense that it is deployed "on the field" in the fight against misinformation or hate speech (among other things) in order to counteract a particular case of misinformation as it arises. Company

[1] The social media companies responded with various types of labels. For example, Twitter used explanatory labeling text such as, "Learn more about US 2020 Election security efforts" with links to informational pages on Twitter, as well as content warning labels such as, "This Tweet is disputed and might be misleading about an election or other civic process" with a link to Twitter's Civic integrity policy. Facebook used content warning interstitials for "false information" for posts claiming election fraud or attempts to intimidate voters; with a "false information" warning on an image, link, or post, users could click through to see the verified fact check sources on election information.

[2] Companies deploy labels for various purposes. For example, Google increased content transparency on YouTube by implementing publisher context labels on videos, which indicate whether a channel is "state sponsored" or is a "public broadcast service" to legitimize reliable information on political news. TikTok was prompted by COVID-19 misinformation to implement widespread labeling on the platform, with Coronavirus information banners on related videos that linked to authoritative health sources. In order to increase friction between misinformation subreddits and Reddit users, the platform implements a "quarantine" on pages—accompanied by a warning label requiring users to explicitly opt-in to view the content in question—that promote conspiracies, hoaxes, and offensive content that violate Community Guidelines, as opposed to labeling individual pieces of content. REDDIT, Quarantined Subreddits, https://www.reddithelp.com/hc/en-us/articles/360043069012 (last visited March 27, 2021).

policies—such as Facebook's "Community Standards" or "The Twitter Rules"—also embody this tactical conception of information quality labeling.[3] The policies are formulated as guidelines regarding the conditions under which the tactic will be employed. Depending on the perceived degree of potential severity or harm, as well as other factors such as the information source (e.g., Twitter has a distinct policy for world leaders), user-generated content may be subject to removal (primarily where physical harm may be involved), algorithmic reduction (making content less visible to other users), or labeling/information treatments, which may surface direct factchecks, more authoritative source information, or further information about the originating source of the content.

However, it is also possible to think of information quality labeling strategically. That is, it is possible to consider information quality labeling as part of an approach to building a healthy informational environment. On this way of considering information labeling, it is not only deployed to combat a particular case of misinformation as it arises, but also to advance the informational quality of the platform overall and the user's ability to effectively navigate the information ecosystem. It is this strategic conception of information labeling that is the focus of this paper. Our aim is to articulate more clearly how and in what sense informational labeling can be used in this way, as well as to identify key ethics and values questions that the platforms ought to consider if they were to do so. The result is an approach for thinking through how to develop a

[3] FACEBOOK, Community Standards, https://www.facebook.com/communitystandards/ (last visited December 23, 2020); TWITTER, The Twitter Rules, https://help.twitter.com/en/rules-and-policies/twitter-rules (last visited December 23, 2020).

proactive and generally beneficial informational quality labeling system.

The keys to thinking about labeling strategically is to consider it from an epistemic perspective and to take as a starting point the "social" dimension of online social networks. These together favor taking a social epistemological[4] approach when thinking strategically about informational quality content labeling, as well as content moderation more generally. That is, platforms should carefully consider how the moderation system improves the epistemic position and relationships of platform users—i.e., their ability to make good judgements about the sources and quality of the information with which they interact on and beyond the platform—while also appropriately respecting sources, seekers, and subjects of information.[5]

In Section One, we provide a review of existing information quality labeling approaches and policies, as well as of the societal and industry context that frames these issues. An emphasis is on how they currently work and associated problems, issues, and challenges. In Section Two, we discuss why a systematic content labeling approach begins with articulating the values and goals of the

[4] "Social epistemology," as we mean the term, is a multidisciplinary field of inquiry that examines the social aspects of thought, rationality, justification, and knowledge and their normative implications. For some core examples of work that aligns well with our general approach to the field see: Alvin I Goldman, *Knowledge in a Social World* (Oxford: Clarendon Press, 1999); Helen E. Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (Princeton University Press, 1990); Miriam Solomon, *Social Empiricism* (Cambridge, Mass.; London: A Bradford Book, 2007); Alvin Goldman and Dennis Whitcomb, *Social Epistemology: Essential Readings* (Oxford University Press US, 2011); Goldman and Cailin O'Connor, "Social Epistemology," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2019 (Metaphysics Research Lab, Stanford University, 2019), https://plato.stanford.edu/archives/fall2019/entries/epistemology-social/.

[5] Kay Mathiesen, *Informational Justice: A Conceptual Framework for Social Justice in Library and Information Services*, 64 LIBRARY TRENDS,198–225 (2015).

moderation regime. In Section Three, we explicate what we mean by taking a social epistemology approach to informational quality content labeling (and to content moderation more generally). We offer new potential measures for defining efficacy and success by content moderation efforts; these proposed measures stand as alternatives to merely limiting and measuring aggregate misinformation spread on platforms. In Section Four, we discuss how normative or ethical considerations can be incorporated into the approach. In Section Five, we conclude by identifying several ways in which the approach could help to inform and improve information quality labeling, as well as to guide further research into such improvements.[6]

DEFINING THE PROBLEM

Complex Mechanics

Content moderation can be defined as the "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[7] Social media companies typically outline their rules in their terms of service and community guidelines, although other policies may apply to content-related decisions. Users are often given some controls such as muting, unfollowing or blocking, as well as organizational options (e.g., by chronology or relevance), which allow for limited local forms of individual moderation.

In general, companies that perform centralized moderation rely on a combination of user reports, or crowdsourced flagging, and automated systems to review content for possible action. Platforms

[6] While our focus here is largely on information quality labeling, the social epistemology approach that we advocate can be applied to other forms—for example, algorithmic interventions and downranking or upranking of content or sources—and targets of content moderation, mutatis mutandis.

[7] James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. TECH. 42 (2015).

with more decentralized or federated content moderation structures and mechanisms, such as Reddit, allow users to perform localized moderation functions within defined communities on the platform.[8] For the purposes of this discussion, the social media platforms using a centralized approach, represented by YouTube, Twitter, Facebook, and Instagram, among others, will be the focus.

Nearly all of the major social platforms spell out guidelines for what is considered violating content and might be subject to removal or other types of actions.[9] Hate speech, violent extremism, harassment, nudity, and self-harm are some of the many categories often subject to heavy moderation and takedowns/removal. Some of this moderation is mandated by long standing laws, such as those relating to copyright violations (e.g., the Digital Millenium Copyright Act, or DMCA),[10] while some newer laws globally, such as Germany's Network Enforcement Act, or NetzDG, are also

[8] For a discussion of the spectrum of content moderation strategies ranging from "industrial" to "artisanal," see: Robyn Caplan, *Content or context moderation?,* DATA & SOCIETY (2018), https://datasociety.net/library/content-or-context-moderation/. There are some recent experiments, such as Twitter's Birdwatch -- a pilot in the US of a new community-driven approach to help address misleading information" -- that allow devolved moderation structures within a platform's larger centralized approach. See TWITTER, "Introducing Birdwatch, a community-based approach to misinformation," https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html (last visited March 20, 2021).

[9] *See, e.g.*, YOUTUBE, Community Guidelines, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#community-guidelines (last visited Dec. 23, 2020); TWITTER, General Guidelines and Policies, https://help.twitter.com/en/rules-and-policies#general-policies (last visited Dec. 23, 2020) ; FACEBOOK, Community Standards, https://www.facebook.com/communitystandards/introduction (last visited Dec. 23, 2020); INSTAGRAM, Community Guidelines, https://help.instagram.com/477434105621119/ (last visited Dec. 23, 2020); TIKTOK, Community Guidelines, https://www.tiktok.com/community-guidelines?lang=en (last visited Dec. 23, 2020).

[10] 17 U.S.C. § 512.

increasingly mandating that social media companies remove defamatory content and hate speech.[11]

False claims, lies, misinformation, misleading statements, and other similar categories generally are not strictly banned by the platforms themselves, unless the speech in question may result in harm of some sort. These non-prohibited categories are the ones increasingly likely to see "softer" information treatments, such as labeling. Labels may be applied that warn users or highlight the disputed nature of content (providing context), and they may rely on and point to external authorities such as media organization factcheckers or governmental agencies as forms of counterspeech. Informational labels may also be accompanied by other social media company actions. For example, on Facebook a labeling treatment when prompted by a fact-check from a third-party may also be accompanied with algorithmic reduction in visibility to other users, or downranking of the content in question and any associated URL across the platform.[12]

Almost every platform's moderation policy leaves room for exceptions based on circumstance. Consider this language from the community guidelines of the social video sharing platform TikTok:

> We recognize that some content that would normally be removed per our Community Guidelines could be in the public interest. Therefore, we may allow exceptions under certain circumstances, such as educational, documentary, scientific, or artistic content, satirical content, content in fictional settings, counterspeech, and content in the public

[11] Heidi Tworek and Paddy Leerssen, *An Analysis of Germany's NetzDG Law*, TRANSATLANTIC HIGH LEVEL WORKING GROUP ON CONTENT MODERATION ONLINE AND FREEDOM OF EXPRESSION SERIES (April 15, 2019) (on file with Annenberg Public Policy Center of the University of Pennsylvania).

[12] *See, e.g.*, FACEBOOK, *Facebook Journalism Project, How Our Fact-Checking Program Works,* (2020) https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works.

> interest that is newsworthy or otherwise enables individual expression on topics of social importance.[13]

Many decisions, in other words, involve judgements based on perceived user intention, social importance, and cultural context. A given piece of questionable content, having been flagged by users or automated systems, typically is sent for a first layer of very cursory human review. Edge cases are then escalated up to content review teams that have increasingly more policy oversight and authority.[14] Given that large platforms have hundreds of millions or billions of users, however, the scale of the content moderation enterprise means that most decisions are the result of either algorithms or the briefest of human review. Indeed, the COVID-19 pandemic and the limitations it placed on office-based work led to many companies such as Twitter, Google/YouTube, and Facebook/Instagram handing over most of their decisions to automated systems.[15] After an initial refusal to release data about enforcement of community guidelines, beginning in 2018 companies such as YouTube, Twitter, Facebook/Instagram started reporting more statistical information about their overall moderation efforts. These reports may include the total volume of content seeing moderation; the prevalence of such categories of content such as hate speech on their platforms; and their rate of preemptive, algorithmic actions before violating content is widely shared.[16]

[13] TIKTOK, Community Guidelines: Introduction (Dec. 2020), https://www.tiktok.com/community-guidelines?lang=kn-IN.

[14] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

[15] Mark Scott and Laura Kayali, *What happened when humans stopped managing social media content*, POLITICO (Oct. 21 2020), available at https://www.politico.eu/article/facebook-content-moderation-automation/.

[16] Daphne Keller and Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, in SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM (Nathaniel

Labeling strategies continue to grow rapidly, in part out of increased pressure from the public, policymakers, and potential regulators, as well as out of a response to extraordinary events such as the COVID-19 pandemic, electoral misinformation, and the violent riots at the U.S Capitol on Jan 6, 2021 that attempted to disrupt certification of the country's election results. For example, many social media companies have created policies that limit attempts to interfere with election procedure (e.g., providing incorrect time of voting), participation (e.g., voter intimidation), or dubious claims relating to fraud.[17] Third-party factcheckers or authoritative sources are sometimes leveraged to add context on a wide variety of these and other kinds of claims. Facebook accompanies various fact-checker findings with ratings such as "False," "Partly False," "Altered," or "Missing Context," while many platforms direct users to more reliable health and election information sources, for example.

Any major technology platform labeling regime faces the problem of scale. Facebook reportedly labeled 180 million messages during the 2020 election season; Twitter stated that it labeled 300,000 tweets during roughly the same period.[18] Both companies have asserted that these labels and warnings resulted in some reduction in the spread of misinformation. Other companies, such as YouTube, took a less targeted approach with respect to the U.S.

Persily and Joshua A. Tucker, eds., 2020). For an example of reporting, see: FACEBOOK, Community Standards Enforcement, (last visited December 23, 2020), https://transparency.facebook.com/community-standards-enforcement.

[17] ELECTION INTEGRITY PARTNERSHIP, Evaluating Platform Election-Related Speech Policies, EiP Policy Analysis (2020), https://www.eipartnership.net/policy-analysis.

[18] Rachel Lerman and Heather Kelly, *Facebook says it labeled 180 million debunked posts ahead of the electio*n, WASH. POST (Nov. 19, 2020), https://www.washingtonpost.com/technology/2020/11/19/facebook-election-warning-labels/; Vijaya Gadde and Kayvon Beykpour, *An update on our work around the 2020 US Elections*, TWITTER CO. BLOG (Nov. 12, 2020), https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html.

2020 election, putting generic labels on a wide variety of election-related content. Taken as a whole, company policies are often incompletely and inconsistently applied, as well as contrary to one another, resulting in content allowable on one platform that may be subject to removal or heavy moderation on another. This is true even in a relatively narrow context, such as electoral integrity, where companies are generally aligned on the goals of free and fair elections but the policy implementation and tactics employed vary widely.[19] This creates an uncertain epistemic environment for users that can undermine trust in a platform's moderation regime, as well as invite accusations of bias, favoritism, and censorship.[20]

Novel Media and Information Ecology

How did we get to such a situation, where the expressions of billions of people around the world are subject to surveillance, filtering and, sometimes, labeling by corporations? Understanding the context that helps explain this historically peculiar situation is crucial to formulating durable strategic solutions.

Major structural shifts in the nature of communications are forcing new discussions about how policies and governance regimes might best preserve public interest considerations for twenty-first century speech environments while also minimizing harms.[21] To be sure, social media companies have themselves created many of the novel problems now requiring remedies by their often-unfettered desire for growth. They have seemingly outstripped their own

[19] ELECTION INTEGRITY PARTNERSHIP, EIP Policy Analysis (2020), https://www.eipartnership.net/policy-analysis.

[20] Emily A. Vogels, Andrew Perrin, and Monica Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints*, PEW RESEARCH CTR. (Aug. 19, 2020), https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

[21] PHILIP M. NAPOLI, SOCIAL MEDIA AND THE PUBLIC INTEREST: MEDIA REGULATION IN THE DISINFORMATION AGE (2019).

abilities to govern their platforms thoroughly and judiciously, a situation fueled by the protections of Section 230 of the U.S. Communications Decency Act, which allows them to avoid liability for user-generated content they host.[22] These structural legal protections have continued to produce negative externalities. Some scholars contend Section 230 is at the core of a wide variety of threats to civil rights and civil liberties — particularly for those without institutional power and groups often targeted for threats and abuse because of race or gender — and thereby constitutes a "discriminatory design" that disadvantages the most vulnerable in society.[23]

As we enter the third decade of the 21st-century, the social media industry stands at a crossroads of sorts. There are tradeoffs between seeking to maximally capture and monetize attention and seeking to elevate high-quality information to minimize harms. The expansive, but essentially ethics-free, nature of Section 230 creates a kind of moral void, according to social media employees, and it drives the need for companies to articulate their own universal "mission" or "central framework," without which company activity lacks clear orientation.[24] Employees within Facebook, for example, have been reportedly split bitterly over how to balance the demands

[22] 47 U.S.C. § 230.

[23] OLIVIER SYLVAIN, *Discriminatory Designs on User Data*, and DANIELLE KEATS CITRON, *Section 230's Challenge to Civil Rights and Civil Liberties*, in THE PERILOUS PUBLIC SQUARE: STRUCTURAL THREATS TO FREE EXPRESSION TODAY (David E. Pozen ed., 2020). While some have argued for removing the Section 230 protections, others have suggested that maintaining them (in some form) could be used as leverage to require platforms to improve content management and moderation practices to promote social goods and values; see Josh Bernoff, *Social media broke America. Here's how to fix it*, BOSTON GLOBE (Dec. 18, 2020), https://www.bostonglobe.com/2020/12/18/opinion/social-media-broke-america-heres-how-fix-it/.

[24] Caplan, *supra* note 8.

for growth with the need to maintain informational quality, civility, and safety on the platform.[25]

Social media platforms have ramped up active content moderation efforts in part to deal with the fallout from a more polarized political environment. The communications spaces they have architected allow both for the expansion of democratic conversation but also the rapid proliferation of hate speech, threats, abuse, and bullying. Millions of people may be exposed to damaging disinformation and misinformation before credible sources can even have the chance to provide opposing views, alternatives, and counterspeech. Algorithms, or computational mechanisms for curation and selection of content, platform designs, and user preferences may also segregate or cocoon people in information silos so that they are not exposed to alternative perspectives or corrective messages. Harms to society may be realized with such scale and speed that traditional safeguards and remedies, namely passively assuming that corrective ideas and accurate information from credible speakers will rise up to compete, seem inadequate, even naive.[26]

The scale of social media, the black-box algorithms that they use, the hyper-personalization of recommendation systems, and the network effects that both lock in the dominance of a select few platforms and enable immense cascades of viral sharing combine to change the fundamental paradigm of speech environments as societies have conventionally understood them. We are quickly

[25] Kevin Roose, Mike Isaac and Sheera Frenkel, *Facebook Struggles to Balance Civility and Growth*, N.Y. TIMES (Nov. 24, 2020), https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

[26] Garrett Morrow, Briony Swire-Thompson, Jessica Polny, Matthew Kopec and John Wihbey, *The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation*, (Dec. 3, 2020) (working paper) (on file with Northeastern University Ethics Institute, SSRN: https://ssrn.com/abstract=).

moving away from the controlling ideas for news and information of the twentieth century, embodied in Justice Holmes's notion, articulated in his famous dissent in *Abrams v United States* (1919), that ultimate goods are produced by the "free trade in ideas" within the "competition of the market."[27] Confusion and misinformation often win the day, with little chance (let alone expectation) for correction or remedy to emerge from the current cacophony of ideas and information. From the prevailing idea of *competition* in the marketplace of ideas, we are moving to a paradigm where individuals' *orientation* and resources for *navigating* the pitfalls of the environment of ideas are becoming paramount.[28] This is why greater regard for the epistemic position of platform users is so important, and why new forms of intermediary interventions—active content moderation approaches—are needed. It is no longer reasonable to believe that the marketplace of ideas will sort the true from the false, the well-informed from the specious, and the well-intentioned from the manipulative.

Substantial policy drift[29]—where old rules remain, but source, platform and consumption patterns continue to be transformed—has taken place across media and communication systems in the United States. This would include Section 230, enacted decades ago, before Facebook, Twitter or YouTube existed. Further, the rise of new technologies has meant that traditional forms of verified news and knowledge have become less central in

[27] Abrams v. United States, 250 US 616, 624 (1919).

[28] This idea of the need for an increased emphasis on user orientation, online cartography, or epistemic positioning has recently been echoed by other theorists. For example, see: WHITNEY PHILLIPS AND RYAN M. MILNER, YOU ARE HERE: A FIELD GUIDE FOR NAVIGATING POLARIZED SPEECH, CONSPIRACY THEORIES, AND OUR POLLUTED MEDIA LANDSCAPE (2020).

[29] For a discussion of the idea of policy drift more broadly, see: J.S. HACKER, P. PIERSON, & K.A. THELEN, *Advances in comparative-historical analysis*, DRIFT AND CONVERSION: HIDDEN FACES OF INSTITUTIONAL CHANGE 180–208 (J. Mahoney, & K. A. Thelen eds., 2015).

terms of public attention, and market structure often no longer sufficiently supports the provision of quality news, shared public knowledge, or exposure to a variety of perspectives.[30] As advertising dollars have moved to online spaces, most have gone to Google and Facebook because of their ability to target consumers based on the vast data they collect, and traditional news media entities have largely lost out.

During this period of drift, few if any policy reforms have been enacted. It should be noted that scholars have long anticipated the need to reexamine the controlling "marketplace of ideas" metaphor, and its policy implications, and contemplated a need to require new forms of disclosure and context to mitigate the pathologies of a more wide-open system of communication.[31] Yet it has taken two decades for many to realize the extent to which the old paradigm has been overturned and novel problems may now require a substantial rethinking of approaches and policy tools.

Moderation and Labeling Challenges

Social media companies are now pouring millions, if not billions, of dollars into content moderation.[32] The new information ecology has created a robust demand for speech regulation, one with radically uncertain rules and few historical precedents with which to help guide the future. Among other anomalies, there is the inherent difficulty of trying to encourage and implement public interest goals and standards on what are in effect private company properties. Further, companies themselves claim First Amendment protections

[30] JOHN P. WIHBEY, THE SOCIAL FACT: NEWS AND KNOWLEDGE IN A NETWORKED WORLD 198-200 (2019).

[31] ALVIN GOLDMAN, KNOWLEDGE IN A SOCIAL WORLD (2002).

[32] Janko Roettger, *Mark Zuckerberg Says Facebook Will Spend More Than $3.7 Billion on Safety, Security in 2019*, NASDAQ (Feb. 5, 2019), https://www.nasdaq.com/articles/mark-zuckerberg-says-facebook-will-spend-more-37-billion-safety-security-2019-2019-02-05.

to defend their right to exercise editorial control of their platform content, although these may be asserted on questionable grounds.[33]

As mentioned, companies have available to them a variety of tools for moderation, including removal and reduction in visibility to users. Until recently, these two approaches were the primary ones employed by companies. But the complexity of regulating political speech, and the ambiguities involved, has forced them to adopt more nimble, "softer" approaches such as warning labels, knowledge panels, source transparency buttons, and other "metadata" instruments, or information about information.[34] While a sizable research literature on platform content moderation has grown as the social web has expanded over the past 15 years, little has been said about content labeling as a comprehensive strategy. Although labeling strategies are highly evolved, and often sophisticated, in other domains such as consumer products, food, pharmaceuticals, and even media- and information-driven spaces such as the entertainment industry, the concept is immature in the social media domain.

There exists a major body of research literature relating to information labeling and disclosure in the context of public regulation and governance,[35] but few have studied how such insights might be operationalized in a social media context. Facebook announced in just 2016 its initial intention to partner with third-party factcheckers, inaugurating a new chapter in the history of online mass content labeling. Even the most comprehensive and recent

[33] LAWFARE, Kyle Langvardt, *Platform Speech Governance and the First Amendment: A User-Centered Approach*, THE DIGITAL SOCIAL CONTRACT: A LAWFARE PAPER SERIES (2020).
[34] Morrow et al., *supra* note 26.
[35] CASS R. SUNSTEIN, TOO MUCH INFORMATION: UNDERSTANDING WHAT YOU DON'T WANT TO KNOW (2020).

scholarly works[36] barely touch on labeling as a standalone, substantive issue. Social media companies are just beginning to take on board the implications of the relevant psychological research literature—for example, the illusory truth effect, the backfire effect, the continued influence effect, and the implied truth effect, among others—and related insights about the correction of information.[37]

There is a strong case to be made that while the companies may have achieved occasional tactical successes in limiting the spread of harmful content and misinformation, in the process they have fostered mistrust in the system by users, undermining their own efforts and inviting objections of political bias, censorship, favoritism, arbitrariness, and amateurism. Media and academic observers frequently note that content moderation decisions by the social media companies are ad hoc and generally reactive, creating what some have called a constant "cycle of shocks and exceptions."[38] Some critics claim that labels are more a form of public relations, and less a substantive attempt to deal with the problem of misinformation.[39]

As reflected by polling data, content moderation strategies have done little to engender trust in social media platforms. As of mid-2020, some three-quarters of Americans believed that platforms intentionally censor certain political viewpoints.[40] On questions specific to the labeling of inaccurate information, there are deep

[36] For example, see one of the seminal monographs in this subfield: TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).
[37] Garrett Morrow et al., *supra* note 26..
[38] Mike Ananny and Tarleton Gillespie, *Public Platforms: Beyond the Cycle of Shocks and Exceptions*, OXFORD UNIVERSITY (2016).
[39] Geoffrey A. Fowler, *Twitter and Facebook warning labels aren't enough to save democracy*, THE WASHINGTON POST, Nov. 9, 2020, https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/.
[40] Vogels et al., *supra* note 22.

partisan divisions, with conservative-leaning respondents overwhelmingly likely to doubt the legitimacy and intentions of social media labeling efforts and liberal respondents split in terms of confidence in the companies to make these decisions.[41] Qualitative research on how users react to content moderation decisions relating to their own posts and accounts suggests deep and persistent public confusion over policies, motives, and reasons for enforcement actions such as content takedowns or account suspensions.[42]

Many of the larger problems with content labeling and content moderation are about more than just questionable tactical judgments or the optics of particular decisions. Rather, the problems are embedded in structural processes and upstream systems—much of which connects to outsourced work of other firms who help with the moderation tasks—set up by the companies, which must execute these policies over user populations of vast scale. The algorithms deployed to assist with this work can miss large amounts of problematic content—particularly when they encounter novel content that does not fit prior patterns of violating content—while also generating false positives. The use of, and claims about, artificial intelligence by the companies should be subject to scrutiny, both on the grounds of ethics/fairness and efficacy/accuracy.[43] The consequences of the more heavy-handed content moderation decisions such as takedowns and removal have seen some amount of careful study, although public understanding remains limited

[41] Vogels, Perrin, and Anderson, *supra* note 20. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

[42] Sarah Myers West, *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms*, 20 NEW MEDIA & SOCIETY 4366-4383 (2018).

[43] Tarleton Gillespie, *Content moderation, AI, and the question of scale*, 7 BIG DATA & SOCIETY (2020), available at doi: 2053951720943234.

because of a lack of full transparency about the platforms' work in this respect.[44]

Despite the limits of algorithms to date, such computational processes are already heavily involved in content labeling regimes, as they are used to track and label, for example, COVID-19- or election-related claims. Increasingly, social media companies are focusing on the authors of misinformation themselves, who tend to be relatively small in number but powerful in their effects on the platform, and their networks that often receive, amplify, and engage with this mis- or dis-information.[45] These two trends—the use of algorithms to scale labeling efforts, and a focus on users who are persistent "bad actors" and their receiving networks— raises the possibility of increased personalization of labeling efforts. There is little public evidence yet of social media companies using algorithms to differentiate labeling strategies for individual content consumers, such that labels seen by one user are not seen by another. But given the social platforms' ability to target and personalize information to users, it would be surprising if more personalized and tailored strategies are not being explored.[46]

Yet the human labor involved in moderation efforts must also remain a key area of critical analysis. As mentioned, teams of moderators are often contract workers employed by outside firms

[44] Daphne Keller And Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM (Persily, Nathaniel, and Joshua A. Tucker, eds., 2020).

[45] Elizabeth Dwoskin, Massive Facebook study on users' doubt in vaccines finds a small group appears to play a big role in pushing the skepticism, THE WASHINGTON POST (March 14, 2021), https://www.washingtonpost.com/technology/2021/03/14/facebook-vaccine-hesistancy-qanon/; Anti-Covid vaccine tweets face five-strikes ban policy, BBC NEWS (March 2, 2021), https://www.bbc.com/news/technology-56252545.

[46] We discuss the dimensions and potential problems associated with deploying a personalized labeling strategy in Section 4 of this paper, where we discuss incorporating normative considerations into content moderation regimes.

working under tight timelines. Overall, content moderation systems are designed with economic and labor constraints that are inadequate to the task of achieving acceptable outcomes. Scholars have explored how outsourced, often under-paid workers help to review content and shown how these systems sometimes result in arbitrary decisions with little remedy.[47] Content moderation teams may need to be significantly expanded and the work function raised to a higher-status role within companies.[48]

However, it should be acknowledged that, as expectations and related regulations for content moderation increase, this may create problems and new complexities. Although this discussion has focused on large, established platforms, there are significant questions about how emerging startups that could challenge incumbents might be expected to resource, at increasingly greater expense, content moderation efforts. If social media are expected to police their platforms with vigilance and consistency from the outset, startup costs may be too high, stifling potential competitors and locking in the advantages of the extant mega-platforms.[49]

In sum, social media companies have been struggling to devise and implement policies on handling misinformation that the public finds generally palatable. In place of consistently enforced policies that are transparent to all parties, the large platforms such as Twitter and Facebook have been handling individual instances of misinformation seemingly piecemeal: downranking some posts, removing others, and labeling or "fact-checking" still others. This

[47] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

[48] Paul M. Barrett, *Who Moderates the Social Media Giants?*, CTR. BUS. N.Y. UNIV. (2020).

[49] Tarleton Gillespie et al., *Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates*, 9 INTERNET POL. REV. 1-29 (2020).

approach has led to social blowback, especially in those cases where black-boxed algorithms downrank or remove posts for stating what might reasonably be interpreted as political or protected speech.

Given the need for these platforms to keep their users happy enough with content moderation policies, it seems likely that the platforms will lean more and more heavily on labeling misinformation, as opposed to removing it or burying it. It appeals as a "middle way" solution for political speech that flags misinformation without fully censoring it, for example, while reliance on third party fact checkers dislocates some of the responsibility from the platforms. It is also, in some respects, the most transparent of the available strategies. It involves providing additional information to users, rather than eliminating or hiding content, and the label and intervention are manifest to users. In contrast, downranking content is a complete black box from the user's perspective and reduces visibility, while censorship is by its very nature opaque.[50]

There is growing sentiment that, as Tarleton Gillespie has advocated, "Platforms should make a radical commitment to turning the data they already have back to [users] in a legible and actionable form, everything they could tell me contextually about why a post is there and how I should assess it."[51] Yet if misinformation is not labeled by these platforms according to a transparent and consistently enforced policy, surely the public will not be much

[50] To be clear, the point here is that labeling is more transparent than alternative strategies, not that labeling is free from any concerns over transparency. See Harrison Mantas, Fact-checkers support Twitter labels, but more than that, they want transparency, POYNTER (May 29, 2020), https://www.poynter.org/fact-checking/2020/fact-checkers-support-twitter-labels-but-more-than-that-they-want-transparency/.

[51] T. GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 199 (2018).

better off. The many problems associated with moderating content on social media platforms suggest that a larger strategic review of the entire problem space is in order. There is a pressing need for a richer and more systematic set of ideas and approaches. This begins with a clear articulation of the goals for the strategy. What, exactly, is the content moderation regime meant to accomplish?

EMBRACING VALUE IN CONTENT MODERATION

What are the Underlying Values and Ultimate Goals of the Moderation Regime?

The considerations discussed above point to the need for a systematic approach to content moderation. In what follows we develop a possible strategic framework for content moderation, including informational quality labeling, that involves: articulating the moderation strategies' goals (and values that underlie them); characterizing the intermediate epistemic aims to accomplish the goals; and identifying ethical considerations (e.g., respect, free speech, equality, justice) that should inform strategies in pursuit of epistemic aims. In this section we argue that developing such an approach requires relinquishing certain myths about platform neutrality.

Social media platforms are designed to be open. (We are here distinguishing network and sharing platforms from more private communication-oriented messaging apps, such as WhatsApp.) The build of the techno-social system is fundamentally oriented toward an increase in users and quantity of information, an increase in connections between users, and facilitation of the movement (or access or sharing) of information across users. What makes them, fundamentally, social media platforms seems to favor a presumption or default in favor of allowing information and smoothing its sharing. At the policy level, the result is an onus or burden of justification on restricting information and spread.[52] It is why the platforms tend to adopt harm-principle oriented policies. This is illustrated in Facebook's policy that highlights two primary aims: Freedom of expression (the default of openness) and avoidance of harm (the consideration that can overcome the presumption of openness).[53] But it also means that content moderation based on information quality is at odds with the design and orientation of not only the companies, but the technologies. Founder and CEO Mark Zuckerberg is quite clear that Facebook the company does not want to be the arbiters of truth[54]; and Facebook the techno-social system is designed in a way that resists evaluating informational quality. Their professed ideal is neutrality.

Social media companies are not the first information institutions to try to take this position. Libraries are information repositories and access systems that have at times embraced the idea of information quality neutrality. Some have argued that the role of libraries should be to make information available, and then leave it up to citizens and patrons to determine what is true or false. On this

[52] Consider the mission statements of two leading companies, which focus on autonomy and lack of barriers. Facebook states: "Founded in 2004, Facebook's mission is to give people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them"; FACEBOOK INVESTOR RELATIONS, Resources: FAQ's (2019) https://investor.fb.com/resources/. Twitter states: "The mission we serve as Twitter, Inc. is to give everyone the power to create and share ideas and information instantly without barriers. Our business and revenue will always follow that mission in ways that improve—and do not detract from—a free and global conversation"; TWITTER INVESTOR RELATIONS, Contact: FAQ's (2021) https://investor.twitterinc.com/contact/faq/.

[53] *See Mark Zuckerberg Stands for Voice and Free Expression*, FACEBOOK NEWSROOM (Oct. 17, 2019), https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/.

[54] Yael Halon, *Zuckerberg knocks Twitter for fact-checking Trump, says private companies shouldn't be 'the arbiter of truth'*, FOX NEWS (May 27, 2020), https://www.foxnews.com/media/facebook-mark-zuckerberg-twitter-fact-checking-trump.

view, labeling for informational quality is seen as a kind of “censorship” because it intervenes between the seeker of information and the source of information. It inserts the librarian's views to influence the seeker’s views. (There are echoes of this in the claim that labeling tweets is a kind of censorship, and that retweeting is not an endorsement.) But library neutrality with respect to information is untenable for at least two interrelated reasons: quantity and organization. There is more information than libraries can make equally available. Therefore, librarians must make decisions about what should be in their holdings, as well as which of their holdings will be made more prominent or easily accessible. The second is that in order to help patrons navigate the vast amount of information, they organize it by category (or directional labeling). They make judgements about what is fiction, what is reference, what is philosophy, what is science, what is research, what is propaganda, and so on. Even if they do not make judgments on the factual accuracy of information, managing the information system requires making judgments about what kind of information each item is.

The analog with social media platforms is clear. The sheer volume of information makes strict informational quality neutrality impossible. It is not possible to just present all the information and let users decide what is true (which, as argued earlier, is also a misconception of the epistemology of social media platforms that belies the “marketplace of ideas” framing of the information environment). And, in fact, the platforms algorithmically curate information all the time. The search engines, recommendation systems, and advertising systems all do this in some form. And how they are oriented is determined by what is valued (or their proxies), such as generating more connections, site clicks, or revenue. Similarly, the user interfaces are designed to organize and present information in a particular format and structure. Users have some discretion over what they see—just as library patrons have discretion over how they navigate a library (or its website)—but there are background design decisions that shape the experience, influence decisions, and define the limits of choice. In libraries they involve such things as subject categorization and search resources. On social media, they are the interfaces, settings, and connection options available to users. There are values related to informational importance and quality, as well as to informational exposure and control, designed in the systems *no matter what*, given the sheer volume and need for organization. So companies cannot claim neutrality with respect to informational quality/importance as a privileged basis for building a content moderation system. It is an old point that values are inseparable from the design of technological systems.[55] But in this context it is worth emphasizing that this applies in particular to values related to quality/importance of information.

We take this to have two implications. First, the current content moderation model is founded on a false presumption that informational neutrality is the starting point and ideal from which moderation deviates and so requires justification. Second, a systematic approach to content moderation—including informational quality labeling—begins with an explicit statement of the goals of and values that underlie the content moderation regime.

Our project here is not to make an argument for particular values or goals that content moderation systems should take. But there are some clear candidates from content moderation policies

[55] ARNOLD PACEY, THE CULTURE OF TECHNOLOGY (1985); Langdon Winner, *Do Artifacts Have Politics?* DAEDALUS 121-136 (1980); Langdon Winner, *Technologies as Forms of Life*, in EPISTEMOLOGY, METHODOLOGY, AND THE SOCIAL SCIENCES (2013).

and recent events, such as: Increasing connectivity while avoiding harms to individuals (these are the ones, as mentioned above, recognized by many of the platforms); maintaining basic social and democratic institutions and practices (or public sphere/decency); reducing racism, sexism, and discriminatory ideologies and practices; amplifying the voice and social impact of people from traditionally marginalized groups; and avoiding collective or social harms. Once the ultimate values or goals of the content moderation system are set, then the question becomes how to accomplish or realize them within the system. Here we believe the social epistemological perspective is crucial. When thinking about realizing the goals, it is important to ask how the features of the system can be modified in order to improve the epistemological position of interacting agents (along with their information environments and their behaviors/judgments) to accomplish these goals or aims.

THE NEED FOR A SOCIAL EPISTEMIC APPROACH

What is a Social Epistemic Approach?

Any systematic and consistent content moderation strategy must first of all be grounded by one or more social values that the strategy aims to promote. But content labeling is essentially an *epistemic* intervention; it is information about information, and so by its very nature, it must promote those social values by making individuals or communities epistemically better off—i.e., by changing their epistemic positions in a way that protects or promotes the ultimate values. As discussed above, when a content moderation regime is overly tactical and reactive it increases confusion, mistrust, and charges of bias—i.e., it does not systematically improve users' epistemic positions. Moreover, social media platform tactics are driven by an unrealistically individualistic understanding of the epistemic contexts and behaviors of their users. Most of the ways in which social media undermines people's epistemic positions are inherently social. The spread of misinformation and fake news are clearly social phenomena, as are the information bubbles and echo chambers users can become trapped within. Such bubbles and chambers tend to erode trust in legitimate sources, limit exposure to alternative views, obscure legitimate expertise, confuse which forms of testimony are evidential, and diminish common knowledge and shared discourse (thereby increasing informational polarization).[56] Any proper content moderation strategy must therefore understand the epistemic concerns in a corresponding way. There are thus two intertwined ways in which the epistemic goals of labeling are social. One is that many of the epistemic outcomes sought are for groups (or for individuals as parts of groups or as they relate to other people)—e.g., avoiding the creation of epistemic bubbles and the erosion of common/shared knowledge. The other is that the social structure of the information system informs what is effective in accomplishing those epistemic outcomes.

This way of thinking in social terms about epistemic interventions is, relatively speaking, a recent advance in the field of epistemology. Besides a few notable exceptions,[57] the study of norms of human thought, rationality, justification, and knowledge prior to the 1980s tended to focus on the sole inquirer, attempting to

[56] Regina Rini, *Fake News and Partisan Epistemology*, 27 KENNEDY INSTITUTE OF ETHICS JOURNAL E–43 (2017); C. Thi Nguyen, *Cognitive Islands and Runaway Echo Chambers: Problems for Epistemic Dependence on Experts*, 197 SYNTHESE 2803–21 (Jul. 1, 2020), available at https://doi.org/10.1007/s11229-018-1692-0; C. Thi Nguyen, *Echo Chambers and Epistemic Bubbles*, 17 EPISTEME 141–61 (June 2020), available at https://doi.org/10.1017/epi.2018.32; Don Fallis and Kay Mathiesen, *Fake News Is Counterfeit News*, 0 INQUIRY 1-20 (Nov. 6, 2019), available at https://doi.org/10.1080/0020174X.2019.1688179.
[57] Here, we have in mind the likes of C.S. Peirce, Émile Durkheim, Thomas Kuhn, and Paul Feyerabend, among others.

build her bank of knowledge from the evidence she had been given by the world itself. Scientists tended to be thought of as isolated individuals, reasoning about nature on their own, and fully outside of any embedded social context. Little attention was given to the fact that most of what humans know they know from the testimony of others, which became an intense topic of debate starting in the 1980s. In the last few decades, epistemologists have recognized that most of what we think, rationally believe, or know for certain traces back to facts about our social circumstances, like whom we talk to, whom we work with, who we take to be experts, how we've been taught to reason by our mentors or society in general, and our informational positions and relationships generally.[58] In other words, we are inherently social inquirers and believers through and through. What we believe, the grounds on which we believe it, and what we know for sure are all features of the particular social epistemic landscape within which we live.

To bring out the limitations of thinking of the epistemic issues in overly individualistic terms, take the following example. In the late summer of 2020, Facebook ramped up its efforts to label posts containing misinformation about COVID-19, examining countless posts and flagging those containing explicitly debunked information. In those cases where posts contained mitigation strategies that conflicted with CDC guidance, context labels were applied, directing users to the CDC's information, on the presumption that users would see the latter as more reliable. The stated aim of these moves was to have fewer individual users exposed to those individual pieces of information. In public statements, Facebook seemed to measure success by the volume of

[58] THE ROUTLEDGE HANDBOOK OF SOCIAL EPISTEMOLOGY (2019), https://doi.org/10.4324/9781315717937.

content that was caught and labeled, and by how much the spread of those particular pieces of misinformation was slowed. But, as watchdog organizations have pointed out,[59] this labeling strategy wasn't able to contain the spread of bogus cures (like Vitamin C[60]), conspiracy theories concerning the origin of the virus (like the 5G conspiracy theory[61]), or anti-vaccination information.[62] What is more, a very large number of platform users seem to still be unable to tell experts from novices, good evidence from weak evidence, or good advice from poor advice on COVID-19 scientific information, and very many have continued to make extremely poor decisions because of it.

Once we move our thinking of content labeling regimes from tactical over to strategic terms, and then ground the strategy in more basic social values, it becomes easier to see that we must think of the epistemic effects of a labeling strategy in social terms as well—e.g., as involving whom to trust, the testimony of others, recognizing expertise, and inferring from the beliefs of others. For example, social media platforms have arguably made it more difficult for members of society to tell who the experts are on a particular topic.[63]

[59] Lukas I. Alpert, *Coronavirus Misinformation Spreads on Facebook, Watchdog Says*, WALL ST. J. (Apr. 21, 2020), https://www.wsj.com/articles/coronavirus-misinformation-spreads-on-facebook-watchdog-says-11587436159.

[60] Reuters Staff, *False Claim: Vitamin C Cures the New Coronavirus*, REUTERS (Apr. 15, 2020), https://www.reuters.com/article/uk-factcheck-coronavirus-vitaminc-idUSKCN21X2PV.

[61] Wasim Ahmed et al., *COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data*, 22 J. MED. INTERNET RES. e19458 (2020), available at https://doi.org/10.2196/19458; Monika Evstatieva, *Anatomy Of A COVID-19 Conspiracy Theory*, NPR (Jul. 10, 2020), https://www.npr.org/2020/07/10/889037310/anatomy-of-a-covid-19-conspiracy-theory.

[62] Talha Burki, *The Online Anti-Vaccine Movement in the Age of COVID-19*, 2 THE LANCET DIGITAL HEALTH e504–5 (Oct. 1, 2020), available at https://doi.org/10.1016/S2589-7500(20)30227-2.

[63] On the epistemology of expertise, see: Sanford C. Goldberg, *Relying on Others: An Essay in Epistemology*, OXFORD UNIV. PRESS (2010); Alvin I. Goldman, *Experts: Which Ones Should You Trust?*, 63 PHIL. AND PHENOMENOLOGICAL

Users seem to have become worse at discerning between a piece of testimony that they ought to trust from one that they ought to discard.[64] This is at least partly because users share information widely with other users without checking the information for accuracy, thus flouting a long-standing norm for making public assertions.[65] Users are often presented with information from an increasingly homogenous set of viewpoints.[66] Those who end up getting fed up with a moderation regime, perhaps because they see it as being politically motivated, might in turn move to a different platform, thus limiting their exposure to an even more homogenous set of views, exacerbating epistemic bubbles and echo chambers.[67]

RESEARCH 85-110 (2001), available at https://doi.org/10.2307/3071090; C. Thi Nguyen, *Cognitive Islands and Runaway Echo Chambers*, *supra* note 51; James Owen Weatherall and Cailin O'Connor, *Endogenous Epistemic Factionalization*, (May 2020) (on file with Department of Logic and Philosophy of Science at the University of California, Irvine).

[64] On the epistemology of testimony see: Miranda Fricker, *Group Testimony? The Making of a Collective Good Informant*, 84 PHIL. AND PHENOMENOLOGICAL RESEARCH 249–276 (2012); DEBORAH TOLLEFSEN, GROUPS AS AGENTS (2015); Jennifer Lackey, *Learning from Words: Testimony as a Source of Knowledge,* OXFORD UNIV. PRESS (2008); Karen Frost-Arnold, *Trustworthiness and Truth: The Epistemic Pitfalls of Internet Accountability*, 11 EPISTEME 63-81 (March 2014), available at https://doi.org/10.1017/epi.2013.43.

[65] On the epistemic norms of assertion, see: Sanford Goldberg, *Assertion: On the Philosophical Significance of Assertoric Speech*, OXFORD UNIV. PRESS (2015); Sanford C. Goldberg, *To the Best of Our Knowledge: Social Expectations and Epistemic Normativity*, OXFORD UNIV. PRESS (2018); John Turri, Truth, Fallibility, and Justification: New Studies in the Norms of Assertion, SYNTHESE 1-12 (2020); Jessica Brown and Herman Cappelen, Assertion: New Philosophical Essays, OXFORD UNIV. PRESS (2011).

[66] On the epistemic concerns raised by homogenous evidence sources, see Kenneth Boyd, *Epistemically Pernicious Groups and the Groupstrapping Problem*, 33 SOCIAL EPISTEMOLOGY 61–73 (Jan. 2, 2019), available at https://doi.org/10.1080/02691728.2018.1551436; Engin Bozdag and Jeroen van den Hoven, *Breaking the Filter Bubble: Democracy and Design*, 17 ETHICS AND INFORMATION TECHNOLOGY 249–65 (Dec. 1, 2015), available at https://doi.org/10.1007/s10676-015-9380-y.

[67] On the epistemology of filter bubbles and echo chambers, see Nguyen, *Echo Chambers and Epistemic Bubbles, supra* at 51; C Thi Nguyen, *Why It's as Hard to Escape an Echo Chamber as It Is to Flee a Cult*, AEON (Apr. 9, 2018), available at https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult.
CAILIN O'CONNOR & JAMES OWEN WEATHERALL, THE MISINFORMATION AGE: HOW FALSE BELIEFS SPREAD (2019).

One consequence of these social-level features of each user's information ecosystem is that many people end up with deeply flawed beliefs both on certain facts about the world that are relevant to their decision making, but also deeply flawed beliefs about whether other people agree with them and share their values. This is evident in some Trump supporters' beliefs that he could not have lost the election without there having been massive fraud, since the vast majority of people that they are exposed to support him and the vast majority of media that they consume support fraud allegations. The deeply social nature of the epistemic situation on social media is central to these kinds of problems.

Re-orienting ourselves toward a more social understanding of the epistemic situation also allows us to see a number of social epistemic benefits that platforms could leverage. For example, social epistemologists have long pointed out that groups of agents can combine to generate epistemic goods that no individual inside the group is capable of (familiar cases are the "wisdom of the crowds" or instances of group knowledge).[68] More recently, network epistemologists have been working on ways to modify social networks in order to increase the likelihood of obtaining certain epistemic goals.[69] And as Neil Levy and Mark Alfano have

[68] JAMES SUROWIECKI, THE WISDOM OF CROWDS: WHY THE MANY ARE SMARTER THAN THE FEW AND HOW COLLECTIVE WISDOM SHAPES BUSINESS, ECONOMIES, SOCIETIES, AND NATIONS (2004); Don Fallis, *Toward an Epistemology of Wikipedia*, 59 J. AM. SOC. INFO. SCI. & TECH. 166 2–74 (2008), available at https://doi.org/10.1002/asi.20870; Alexander Bird, *Social Knowing: The Social Sense of 'Scientific Knowledge*, 24 PHILOSOPHICAL PERSPECTIVES 23–56 (Dec. 1, 2010), available at https://doi.org/10.1111/j.1520-8583.2010.00184.x; Søren Harnow Klausen, *Group Knowledge: A Real-World Approach*, 192 SYNTHESE 813–39 (Nov. 20, 2014), available at https://doi.org/10.1007/s11229-014-0589-9.

[69] Conor Mayo-Wilson, Kevin Zollman, and David Danks, *Wisdom of Crowds versus Groupthink: Learning in Groups and in Isolation*, 42 INTERNAT'L J. GAME THEORY 695–723 (Aug. 1, 2013), available at https://doi.org/10.1007/s00182-012-0329-7; Kevin J. S. Zollman, *The Epistemic Benefit of Transient Diversity*, 72 ERKENNTNIS 17–35 (Oct. 22, 2009), available at https://doi.org/10.1007/s10670-009-9194-6.

convincingly argued, human history is filled with advances in knowledge that seem to be spawned by epistemically problematic behavior if we were to look just at individual inquirers.[70] A more social understanding of the problem might also suggest alternative labeling or context-providing strategies, such as reliability ratings for sharers or sources of information (based on their history) or designing systems so that sharing (or retweeting) requires users to be clear about whether they are actually endorsing what they share.[71] Our suggestion here is that if a content labeling strategy were to respect the deeply social aspects of the epistemic situation with which it is wrapped up, it would not only be able to avoid the various pitfalls of a more individualistic approach but may also be able to generate epistemic benefits that would have been missed by an individualistic, tactical approach. Or, to put it another way, a strategic social epistemology approach is not focused on individual pieces of information or even individual judgments or beliefs about them. It concerns the epistemic relationships and situations of the users collectively.

Case Study for the Social Epistemic Approach: *Trump v. Twitter*

In order to gain a better grasp of what it means and why it is important to take a social epistemology perspective and approach to content moderation, consider again the example of Twitter labeling as "disputed" and potentially "misleading" President Trump's tweets claiming that he had really won the 2020 presidential election and that there had been widespread voter fraud to steal it from him.

[70] Neil Levy and Mark Alfano, *Knowledge From Vice: Deeply Social Epistemology*, 129 MIND 887–915 (Jul. 1, 2020), https://doi.org/10.1093/mind/fzz017.

[71] Rini, *supra* note 51; Rachel Sterken, Jessica Pepp, and Eliot Michaelson, *On Retweeting* (2019) (Manuscript, forthcoming).

Twitter suggested that its 2020 election-related labels limited user sharing of misinformation, "due in part to a prompt that warned people prior to sharing."[72] Here one can see that Twitter is suggesting that the labels were efficacious in reducing the spread of the false claims.[73]

Even assuming this is true, that the labels significantly reduced retweeting and so reduced the spread (or views) of the president's misinformation, Twitter's rationale and approach nevertheless amount to what we have referred to as a very individualist and tactical way of thinking about the misinformation problem and what counts as a solution. From a social epistemology perspective, the question is not how many people on the platform were exposed to the tweet. It is how the labeling changed their epistemic position — and not just about their credence with respect to that particular piece of information. Here are some questions to ask: Did people who were exposed to not just this labelled tweet but a series of them, begin to think differently about how reliable the President was about election information? If so, was it an improvement with respect to their ability to discern misinformation from reliable information? If labels do not change how people structure their information environment, improve their ability to discern misinformation, and lead them to trust more reliable (and mistrust less reliable) sources, then the fact that labelled tweets were

[72] TWITTER COMPANY BLOG, An Update on Our Work around the 2020 US Elections, Nov. 12, 2020, https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html (last visited Dec. 22, 2020).

[73] Paul Mena, *Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook*, 12 POL. & INTERNET 165–83 (2020), available at https://doi.org/10.1002/poi3.214; Geoffrey A. Fowler, *Twitter and Facebook Warning Labels Aren't Enough to Save Democracy*, THE WASHINGTON POST (Nov. 9, 2020), https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/.

viewed less frequently than they would have otherwise been is not an epistemic success. In fact, if persistent, robust labeling leads people to become more discriminating in a way that improves their ability to identify misinformation, then reducing the exposure to labeled misinformation is actually not an epistemic good. Or, to put it another way, the challenge from a social epistemology perspective is not "how to make truth travel faster than lies,"[74] it is how to improve people's ability to distinguish truth from lies in a socially networked informational context.

Major platforms—e.g., Twitter, Facebook, Youtube—ultimately suspended Trump's accounts, on the basis of inciting violence, in the wake of the January 6, 2021, attack on the U.S. Capitol that he helped to foment. This is clear evidence of the failure of their content moderation approaches. The labeling tactics they employed to combat misinformation around the election were ineffectual, and their broader content moderation policies (including the recommendation systems and hyper-personalization they use) fostered radicalization and the growth of white nationalist extremist groups that were central to the riots. The entire (and ongoing) situation demonstrates the importance of thinking about content moderation from a long-term strategic social epistemology perspective. By the time the platforms began to tactically label Trump's posts, the epistemic damage had already been done. Those who were sympathetic to him trusted his claims—even in the absence of supporting evidence and the presence of countervailing evidence, and even with extensive reliable expert testimony and confirmation from numerous vetting and auditing processes. They

[74] Geoffrey A. Fowler, *Twitter and Facebook Warning Labels Aren't Enough to Save Democracy*, THE WASHINGTON POST (Nov. 9, 2020), https://www.washingtonpost.com/technology/2020/11/09/facebook-twitter-election-misinformation-labels/.

were situated in epistemic bubbles and echo chambers that continually reinforced their views. They disbelieved platform labels and distrusted fact-checkers and independent news organizations. By not having had a long-term, value-grounded, consistent, clearly articulated labeling strategy (and broader moderation strategy), the social epistemic situation was such that ad hoc tactical labelling (indeed, any tactical intervention) was bound to fail.

In fact, from a social epistemological perspective, banning Trump from the platforms appears to have had limited effect thus far. His core supporters' epistemic situation has not significantly improved, and the bans have reinforced many of their epistemic priors about bias, conspiracy, and who to trust. Again, the moderation problem is not best understood by focusing on individual posts or numbers of views, but by the sort of epistemic contexts and relationships that platform designs, policies and interventions have helped to build (both on and beyond the platforms). Views of Trump's posts on the platforms that have banned him have gone to zero, but the more important question in evaluating the ban's effectiveness is how has this impacted the problematic epistemic environment that enables conspiracy theories, election misinformation, and hate groups to prosper. Still more important is how to begin to strategically build a content moderation and labeling regime over the long-term that will create a better social epistemic environment and enable effective tactical interventions when future need arises.[75]

[75] Of course, one way to begin to do this is to audit how the current problematic epistemic environment arose, such as the rabbit-holes toward radical content that recommendation systems often create, the ambiguities of meaning and responsibility around retweeting, the inconsistency of the 'newsworthy' exemption, the design features that foster epistemic bubbles and epistemic polarization, the hyper and unrelenting personalization of content that erodes shared knowledge, the absence of a consistent, intelligible and research-based labeling strategy, and so on.

Slowing the spread of lies relative to truth on this or that platform might be a means to accomplishing the goal of improving a user's ability to distinguish truth from lies online, but a lot would depend on the details. If people's epistemic position is not improved, and they instead jump to a different platform with even less content moderation, then that is not success. If suppression or other attempts to mitigate lead people to strengthen their convictions about conspiracies and misinformation (as one might expect due to the self-sealing nature of conspiracy theorizing),[76] then that is not success.[77] It is social epistemic success that is needed, and that might mean more robust labels with links to correct sources are preferable to suppression.[78] Or, to put this another way, it is not the spread of lies that is itself the epistemic problem, it is the way in which those lies lead people to believe more false and fewer true things in the future on the basis of the relational aspects of networked information exposure, and then the costs (personal and social) associated with that.[79]

[76] Cass R Sunstein and Adrian Vermeule, *Conspiracy Theories: Causes and Cures*, 17 J. POLITICAL PHILOSOPHY 202–227 (2009).

[77] Stephan Lewandowsky, KH Ullrich, Colleen M. Ecker, Norbert Schwarz Seifert, and John Cook. *Misinformation and Its Correction: Continued Influence and Successful Debiasing*. 13 PSYCHOLOGICAL SCIENCE 3 106-131 (2012).

[78] Morrow et al., *supra* note 26, summarizes the extant research literature and concludes: "a label should directly refute the misinformation, provide an alternative explanation if available, and provide a detailed explanation with regard to why it is false. The label may be more effective if it comes from someone ideologically aligned with the recipient and includes graphical elements, or other aesthetic elements in line with the affordances and usage practices of the platform's content" ; Briony Swire-Thompson and David Lazer, *Public health and online misinformation: challenges and recommendations*, 41 ANN. REV. PUB. HEALTH 433-451 (2020); Briony Swire-Thompson, Joseph DeGutis, and David Lazer, *Searching for the backfire effect: Measurement and design considerations* (2020) (submitted for publication (on file with PsyArXiv, available at doi:10.31234/osf.io/ba2kc).

[79] It's not just that more people believe that Trump won, but also that fewer people are as confident as they ought to be that Biden did, which imposes costs on the democratic process. These include: costs to news agencies which might need to trim important content in order to spend time debunking the misinformation, possibly causing a drop in viewership; costs to the public officials who are

What is an alternative, social epistemological measure to misinformation spread of whether labeling strategies are effective? We offer a number, which are not intended to be exhaustive:

1. A change in the ratio of posts containing verifiable information to those containing misinformation on a platform.[80]
2. Whether users become better judges of genuine expertise on the topics (as evidenced through their linking, liking, or visiting behavior).
3. Whether users curate their information environment differently with respect to who they follow, unfollow, or block.
4. Whether users are exposed to (or seek out) a wider range of viewpoints on those topics that are still under legitimate dispute.
5. How users alter their sharing behavior (e.g., retweeting) with respect to misinformation (e.g., do they increasingly identify it as such?).

What measure is appropriate to use will depend in part on what social values the content labeling strategy is designed to promote. For example, if the social values require that individuals have accurate beliefs about some subset of factual matters, then the relevant measure will certainly have to take into account whether users of the platform end up with more accurate beliefs on that subject matter as they engage with the platform. On the other hand, if the social values require that individuals take seriously the beliefs

targeted by false rumors or even full-blown conspiracy theories; and costs to overall standards of social discourse and civic engagement, as well as democratic processes and values.

[80] There have been various calls to change the verifiable information-misinformation ratio through much greater knowledge curation by the social media companies. For example, see: Hanaa Tameez, *Beyond "Yellow Banners on Websites": How to Restore Moral and Technical Order in a Time of Misinformation*, NIEMAN JOURNALISM LAB (December 1, 2020), https://www.niemanlab.org/2020/12/beyond-yellow-banners-on-websites-how-to-restore-moral-and-technical-order-in-a-time-of-misinformation/.

and viewpoints of users from opposing sides of the political spectrum, whether users end up having inaccurate beliefs about the former subject matter might be less relevant. In short, which epistemic goals a content labeling strategy ought to promote will depend on the ultimate social purpose the strategy was designed to accomplish, and that will in turn inform what measures should be used to evaluate candidate strategies.[81] This process is largely an empirical matter. It is an empirical question whether this or that content labeling strategy really does make the resulting information ecosystem better or worse on that chosen metric.

To be clear, the empirical studies to distinguish which is the epistemically preferable strategy and measures are nascent,[82] and therefore we are not in a position to settle these issues (in addition to the fact that we are not here endorsing any particular social goals). The point is that how to understand the problem and what constitutes success with respect to addressing it depends on the way it is analyzed, and that insights from a social epistemological perspective offer crucial perspectives on the problem. (Also, to be clear, our point is not that it is the *only* useful one, nor is it to deny that reducing the spread of misinformation is often also important.) We are not the first to make the point that a social epistemology perspective should be central to analysis of and responses to online

[81] It is important to note that these empirical questions also need to account for the international reach of content moderation policies. One might predict that large corporations stationed in a certain nation, such as Facebook with America, might focus on the effects their moderation regime has on users hailing from the same nation. But, as is now well accepted, psychological effects often differ from nation to nation, and thus it would be a mistake to base policies with international reach on studies that lack it. See, e.g. Joseph Henrich, Steven J. Heine, and Ara Norenzayan, *The Weirdest People in the World?*, 33 BEHAVIORAL AND BRAIN SCIENCES 61–83 (June 2010).

[82] Morrow et al., *supra* note 26.

misinformation.[83] But our hope is that the preceding discussion elucidates what it means to approach informational quality content moderation from a social epistemology perspective and how it provides a useful perspective for analyzing the problem of content moderation and developing and evaluating candidate approaches to addressing it.

INCORPORATING NORMATIVE CONSIDERATIONS INTO A CONTENT MODERATION REGIME

The strategic approach to informational labeling that we have advocated begins with clearly articulating the moderation regime's goals (what it is meant to accomplish) and guiding values (why it is meant to do so). Once these are articulated, then it is possible to inquire (from a social epistemology perspective) how the epistemic position of users could be improved through informational labeling to accomplish those goals. Content moderation strategies and policies can then be developed and assessed (using appropriate measures) for realizing those epistemic improvements.

However, there are considerations that must inform evaluation of candidate labeling policies and strategies, which go beyond their efficacy in improving users' social epistemic position according to well defined metrics. Some of these considerations are practical or concern feasibility. Whatever the strategy is, it must be scalable and timely, for example. Given the volume of content to review, this suggests that there will be an automated or algorithmic component. As discussed earlier, there are significant unanswered questions (which we are not addressing here) about how to do this effectively and responsibly. Companies' impulses to try to make the

[83] Rini, *supra* note 51; Nguyen, *Echo Chambers and Epistemic Bubbles, supra* note 51; Fallis and Mathiesen, *Fake News Is Counterfeit News*; Sterken, Pepp, and Michaelson, *supra* note 65.

moderation process more efficient and less susceptible to human bias and error—fueled by technical advances in machine learning/artificial intelligence (ML/AI) and natural language processing (NLP), as well as computer vision—will make the ever-increasing use of automation tempting to the platform companies. However, scholars continue to have concerns that, in fact, AI will amplify existing biases and perpetuate systemic injustices, and that deep-learning algorithms and the like are far less effective than technologists would claim in their ability to grapple with nuanced, often novel, content.[84]

But other considerations are less practical and more normative. A strategy might be epistemically beneficial but nevertheless be contrary to legal or ethical norms. Imagine that a platform implemented a system that downranked (or negatively labeled) posts by people who subscribe to or are regularly exposed to information from some particular media ecosystem because it (the algorithmic system) learned that people who are thus connected tend to share scientific and election misinformation at a high rate. Imagine, further, that the media ecosystem has a particular political orientation. Even if the moderation system was not intentionally designed to slow information spread from individuals who subscribe

[84] Ifeoma Ajunwa, The paradox of automation as anti-bias intervention, 41 CARDOZO LAW REVIEW 54 (2020); Robert Gorwa, Reuben Binns, and Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, 3 BIG DATA & SOCIETY 1–15 (2020). As alluded to earlier (Section 1C), one potential application of these sorts of algorithms in the context of labeling could be to use them to try to predict what sorts of informational quality labels are likely to be most effective for different groups of people in different contexts. That is, it might be possible to employ the sort of algorithmic, data-driven personalization currently used to optimize for engagement with advertisements and products to optimize for engagement with information quality labels and corrective information (e.g. fact-checkers and authoritative sources) as part of an attempt to accomplish positive epistemic outcomes. However, it is important to recognize that concerns over ML/AI generated informational biases could arise if labeling regimes are algorithmically personalized and tailored to each particular information consumer.

or are otherwise exposed to that ecosystem—and even assuming it effectively accomplished the goal of slowing the spread of uncontextualized scientific and election misinformation that erodes people's epistemic position—there could nevertheless be concerns on other grounds. One concern might be on grounds of bias, if the moderation system slowed not only the targeted misinformation but also other (non-targeted) information or views from those sources and users. Another concern might be that it does not treat users on the basis of their own behavior, but instead makes judgments on the basis of informational relationships. It epistemically downgrades users (it reduces their ability to share information) whether or not they themselves are purveyors of misinformation, based on the algorithmic determination that they are the type of user (based on their informational associations) that is likely to do so. In some (but not all) contexts, this sort of judging based on grouping is problematic,[85] and it may be so when it involves restricting or limiting speech. For this reason, such a strategy—one that labels on the basis of informational association—might in some contexts be less desirable than one that is oriented around users' own information behaviors.

There are, in fact, a host of normative considerations relevant to evaluating candidate strategies. Concerns about bias, fairness, censorship, respect, autonomy, rights, accessibility, and equality need to be taken into account. A strategy that is epistemically effective in general or over a large population of users might treat some groups of users differently—for example, labeling their posts at a higher rate or having a higher rate of mislabels—and

[85] Daniel Susser, *Predictive Policing and the Ethics of Preemption* (Ben Jones & Eduardo Mendieta eds.) (forthcoming, on file with NYU Press), available at https://philpapers.org/rec/SUSPPA.

so be problematic.[86] It might not respect the autonomy of users or treat them as individuals in contexts when doing so is required. It might marginalize some persons' or groups' information or perspectives without warrant. It might be comparatively ineffective at reducing misinformation about particular groups of people. It might place undue burdens or costs on some people or groups (e.g. with excessive exposure to corrections or labels). And so on.[87]

The aim here is not to articulate the full range of normative considerations, let alone substantively specify them to the extent that they could be used to evaluate concrete strategies. That is well beyond the scope of this paper. However, we do want to emphasize, following Kay Mathiesen's work on informational justice, that when conducting an ethical analysis to identify potential normative considerations regarding the impacts of information systems on people and groups, it is necessary to take into account the seekers of information (i.e., the content consumers), the sources of information (e.g., the posters and sharers), and the subjects of information (i.e.,

[86] A number of moderation efforts have turned out to be biased against groups whose information behaviors and speech deviates from those on which algorithms are trained or standards developed. This is an area where content moderation is subject to the same sorts of algorithmic bias concerns, such as unrepresentative training data and disparate impacts, that arise in other contexts, such as criminal justice, education, and social services. A rich critical literature has documented these problems across numerous domains. See: Angwin, Julia, and Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, PROPUBLICA (2017); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Sloane, Mona, and Emanuel Moss, *AI's social sciences deficit*, 1 NATURE MACHINE INTELLIGENCE 330-331 (2019); RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (2019); MEREDITH BROUSSARD, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (2018).

[87] For in depth treatments of related concerns over epistemic distributive justice see Kurtulmus, Faik and Gurol Irzik. 2017. "Justice in the Distribution of Knowledge." Episteme 14: 129-46.; Fallis, Don. 2007. "Epistemic Value Theory and the Digital Divide." Information Technology and Social Justice, eds. E. Rooksby and J. Weckert, Idea Group, pp. 29-46.

individuals that posts or claims are about);[88] and that here, too, a social epistemology perspective is helpful because the way in which content moderation works is by altering informational relationships and epistemic positions.

As discussed above, it may be morally problematic if a content labeling strategy treated content consumers from certain groups substantially differently than others. There are a number of different ways in which such strategies do wrong to those seekers who are treated worse than others, particularly if it is members of a protected and typically marginalized group who are made epistemically worse off or if legitimate political/public speech or dissent is suppressed or marginalized.[89] Strategies arguably can also do wrong to information seekers by making members of certain other groups disproportionately *better-off* (even if no users are made straightforwardly *worse-off*). For example, if a content labelling strategy leaves less educated individuals in roughly the same epistemic situation, while drastically improving the epistemic position of those with more education, this also seems, at least *prima facie*, to be of concern. If the platform has access to a slightly non-optimal strategy that also raises less educated seekers, then that may be a strong enough consideration to favor adopting the less-optimal option. In short, many of the same kinds of concerns over

[88] Johannes J. Britz, Making the Global Information Society Good: A Social Justice Perspective on the Ethical Dimensions of the Global Information Society, 59 JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 1171–83 (2008), available at https://doi.org/10.1002/asi.20848; Kay Mathiesen, Access to Information as a Human Right (2008) available at SSRN 1264666; Kay Mathiesen, Informational Justice: A Conceptual Framework for Social Justice in Library and Information Services, 64 LIBRARY TRENDS 198–225 (2015), available at https://doi.org/10.1353/lib.2015.0044; Kay Mathiesen, The Human Right to Internet Access: A Philosophical Defense, 18 THE INTERNATIONAL REVIEW OF INFORMATION ETHICS 9–22 (December 1, 2012).

[89] For example, imagine how different the world would be, from a social justice perspective, if content moderation regimes had inadvertently suppressed the Arab Spring.

distributional justice can also apply to distributions of epistemic goods for information seekers. Distributive justice and fairness are of course not the only normative consideration regarding seekers of information—concerns related to seeker autonomy are also relevant, for example—but it illustrates the need to respect and consider content consumers (and not just content sources) in evaluating candidate strategies and policies.

Respecting the sources of information, which in the social media context tend to be those creating or sharing information with their posts, also generates normative considerations relevant to evaluating content labeling policies and strategies. Perhaps the most commonly discussed instance of this involves censorship and speech rights. These are most often framed as concerns about the treatment of informational sources. (Censorship can also be framed in terms of information access from the perspective of information seekers.) However, taking an epistemic approach reveals other normative considerations. One such consideration, which has been highlighted in the social epistemology literature, stems from concerns over what Miranda Fricker calls "testimonial injustice" (although the general idea was raised in much earlier work by feminist women of color).[90] The large and quickly growing literature on this kind of epistemic injustice documents the many ways in which people from marginalized groups—e.g., women, non-binary persons, people of color, children, overweight people, the elderly, people with disabilities, etc.—are often treated differently, and are very often disadvantaged, as sources of information. The root issue is that individuals who belong to these groups tend to be treated by others

[90] MIRANDA FRICKER, EPISTEMIC INJUSTICE: POWER AND THE ETHICS OF KNOWING (1st ed. 2009); Rachel McKinnon, *Epistemic Injustice*, 11 PHILOSOPHY COMPASS 437–46 (Aug. 1, 2016), available at https://doi.org/10.1111/phc3.12336.

as much less reliable as sources of information than they, in fact, are. Based on this literature, it seems likely that content labels could have differential effects depending on the demographic characteristics of the sources of the information. For example, a corrective content label applied to a piece of misinformation posted by a wealthy, adult, white male might generally be disregarded, while a content label applied to a piece of misinformation posted by a younger, non-wealthy woman of color might cause users to discredit that information at a higher rate. In short, some content labeling strategies may exacerbate forms of epistemic injustice that have already been well documented, and this should be considered when evaluating which strategies platforms should use. Lastly, if a labeling strategy itself treats sources of information belonging to protected demographic categories in substantially detrimental ways, as some have argued has already occurred with other content moderation strategies along racial lines,[91] this would obviously also raise moral concerns.

Lastly, there are also legitimate normative concerns that are related to how subjects of information are affected by a content labeling strategy. For example, if falsehoods posted about White subjects of stories are labeled more frequently than falsehoods posted about Black subjects of stories, then the approach is biased. As mentioned, it is well documented that algorithmic systems can be biased in numerous ways and for numerous reasons, and this applies as well to labeling or moderation algorithms. Moreover, as discussed earlier, there is often a human element to many content

[91] Aaron Sankin, *How activists of color lose battles against Facebook's moderator army*, REVEAL, August 17, 2017, https://www.revealnews.org/article/how-activists-of-color-lose-battles-against-facebooks- moderator-army/; Sam Levin, *Civil rights groups urge Facebook to fix 'racially Biased' moderation system*, THE GUARDIAN, January 18, 2017, https://www.theguardian.com/technology/2017/jan/18/facebook- moderation-racial-bias-black-lives-matter

moderation efforts. The fact that individuals tend to harbor unconscious biases against members of certain groups is well established, and such biases will creep into content moderation efforts. When there are biases in labeling—algorithmic and/or human—they generate epistemic biases. Some people, perspectives, or information are epistemically disadvantaged within the system, for example by being misrepresented or by limiting their ability to represent themselves (and so compromising their autonomy).

These are just sketches of some normative considerations that arise when evaluating content moderation, and information labeling in particular, from a social epistemology and informational justice perspective. They are by no means exhaustive. Moreover, as indicated above, our aim here is to present an approach for strategic use of content labeling—one oriented around social epistemology—and indicate some of the ways in which that approach can be helpful for elucidating the challenge and developing strategies and policies for addressing it. There are other critical perspectives that are useful and other normative considerations that are relevant in addition to those discussed here. The crucial point is that those who wish to develop a robust, systematic content moderation strategy will have to take into account normative and value considerations at several levels. One is in defining the goals of the regime and the values that underlie them. Many of these will be social goods and values, in addition to the value of individual expression and the avoidance of harms. Another is in evaluating the impacts of candidate strategies to accomplish those goals on individuals and groups, including those individuals and groups living in lands very far removed from the developers or implementers of the strategies. In this section we have tried to motivate the importance of analyzing these impacts from the perspective of respect for seekers, subjects, and sources of the information being moderated, as well as the importance of including a social epistemology perspective.

CONCLUSION

We have tried to elucidate a strategic way of framing the problem of online content moderation, one that is grounded in analyzing the problem through the lens of social epistemology. The framework we are proposing involves: Identifying and articulating the ultimate goals (and the values that underlie them) to be accomplished by the moderation strategy; determining what epistemic impacts (changes to information context and agents capacity to navigate it) are needed to accomplish those goals; developing normatively informed strategies and tools to accomplish those epistemic aims (and evaluating them accordingly). We have highlighted several ways in which taking this approach might inform, and in some cases improve, content moderation in general, and informational quality labeling in particular.

Consistency and coherence: The largely reactive and piecemeal approach to content moderation policy and practice is an underlying cause of a number of difficulties in content moderation. Charges of bias and favoritism arise. Moderation activities appear ad hoc. There is overall a lack of coherence in the discourse and practice around content moderation. It is difficult to argue tactics—what works, what does not, what is acceptable—when the end goal is not at all clear, or is narrowly tailored to stopping misinformation spread. The framework we propose begins with clearly articulating the ultimate goals (and the values that justify them) of the moderation regime. This benefit is not particular to the framework we have proposed here; it is a general benefit to any clearly articulated, longitudinal and systematic approach. Of course, adopting a clear strategic framework does not ensure consistency in application, but it is difficult to imagine consistency without one (i.e., it's necessary, not sufficient).

Understanding harms. There is widespread agreement that current moderation practices are inadequate. But in order to develop solutions, it is important to be able to characterize more precisely how they are inadequate. As discussed above, individualistic harm-based analyses are insufficient. The types of harms that misinformation contributes to are collective and social as well. Moreover, the ways in which those harms are realized is often through eroding the social epistemic position of users with respect to evaluating sources of information, what information and sources to trust, and the diversity of informational sources and perspectives to which they are exposed, for example. And because platform users living in different cultural contexts will often have very different social epistemic contexts as well, harms are also likely to differ across national or cultural boundaries. A social epistemic analysis of and approach to content moderation therefore helps to more fully characterize the content moderation problem and the associated harms/wrongs involved.

Defining success: As discussed above, it is crucial to have a clear account of what counts as success in a labeling strategy (or any content moderation strategy). A social epistemology approach favors thinking about success in terms of epistemic impacts systematically, rather than in terms of exposures or access. The question is not how many people see something, but how they are seeing it, and the ways in which it leads changes to their epistemic position with respect to things such as information exposure, whom they trust, what they take as authoritative, and the diversity of informational sources/perspectives.

Measuring success: Measures of success should reflect the definition of success. Are users better constructing their epistemic space as defined by the success criteria? Are their information behaviors (sharing, endorsing, posting) improving in response to the labels as defined by the success criteria? A feature of social networks is that users are co-curators of their, and their networks', information exposure. So it should be possible to measure changes in their epistemic situation in response to persistent labeling by looking at such things as changes in the frequency with which they share labeled information, the frequency with which they engage in endorsing behaviors for labeled information, whether they begin dropping or reducing connections to users who are persistently negatively labeled, and whether they look for or explore alternative or more diverse informational sources. What works (like what ultimate values and normative considerations are most salient) may differ by cultural context.

Needed platform data and experimental research. The experiences of the major platforms in 2020 relating to COVID-19 and the U.S. election have produced extraordinary data about content labeling that, so far, is only accessible to the platform companies. Measuring success, and thereby assessing efficacy of information quality labeling and other moderation strategies according to a social epistemology (or any other) strategic approach, is only possible if researchers have access to the data. How those millions of content labels affected user behavior, both immediately and over longer periods, is a rich potential area of inquiry, including from a social psychology perspective. Those data might point to informational interventions that modified behavior in positive ways, suggesting boosts that provide epistemic positioning for users. Platform data about the use of fact-checking more generally and its consequences remain inaccessible, and the companies need to share much more of this in order to help both researchers and factcheckers improve outcomes.[92] For example, it would be very useful to conduct experiments on platforms that vary approaches, such as using more graphical information and providing more detail about sources. Importantly, this could help researchers better understand how to tailor labels to help put lower-literacy and/or lower-knowledge users in a better epistemic position, or how to tailor them for different informational and cultural contexts. (This research would be analogical to research on content label designs and efficacy for nutritional and other food labeling.) It is also crucial to determine, in the context of a labeling practice, how users respond to unlabeled information and sources—e.g., Do they presume reliability in the absence of a negative label?—as what matters most from a social epistemology perspective is not how users interact with labeled content, but how labeling practices impact users' overall epistemic position. At the end of the day, any public policy changes, such as modifications to Section 230, should take into account what responsible content moderation looks like when it does more than just limit the spread of misinformation, but rather improves the epistemic environment for a democratic citizenry very much in need of better orientation.

Innovating new strategies. Taking a social epistemological approach can help foster innovative thinking on possible interventions. Instead of asking how to slow the spread of misinformation or improve individual critical thinking skills, it

[92] The lack of data access from companies remains a major obstacle to independent empirical research of many kinds. For a major statement on this issue from many leading researchers in the field, see: I. Pasquetto, B. Swire-Thompson, M.A. Amazeen, F. Benevenuto, N.M. Brashier, R.M. Bond, L.C. Bozarth, C. Budak, U.K.H. Ecker, L.K. Fazio, E. Ferrara, A.J. Flanagin, A. Flammini, D. Freelon, N. Grinberg, R. Hertwig, K.H. Jamieson, K. Joseph, J.J. Jones, … and K.C. Yang, *Tackling misinformation: What researchers could do with social media data*,1 HARVARD KENNEDY SCHOOL: MISINFORMATION REVIEW (2020).

invites exploring strategies that could improve epistemic positions and relationships of users. For example, a social epistemology perspective has led to suggestions around labeling sources and sharers of information (rather than just pieces of information),[93] as well as norm engineering around retweeting.[94] It might also inform thinking about how to design user co-curation options to enable or nudge them toward better (as understood through the epistemic aims) information curation and sharing, for example by inviting them (and making it easy) to unfollow or block sources or sharers of persistently labeled misinformation.

Situating ethical considerations. There is widespread recognition that ethical considerations are relevant to content moderation. However, it is often unclear what, precisely, the ethical considerations are and how they ought to figure into decisions regarding content moderation. The framework offered here begins to explicate both of these. On the framework, ethical considerations are relevant to establishing overarching content moderation goals, as well as to evaluating candidate content moderation strategies. The informational justice approach helps to identify a fuller range of ethical considerations that are relevant by encouraging evaluation of policies and practices from multiple perspectives, including sources, seekers, and subjects of information.

Again, our goal here has been to elucidate an approach for analyzing and responding to the content moderation problem. We have argued that an ethically informed social epistemology approach can provide a helpful perspective on informational labeling and content moderation more generally. In some senses, this has been an exercise in ideal theorizing about content

[93] Rini, *supra* note 51.

[94] Sterken, Pepp, and Michaelson, *supra* note 65.

moderation. We have not addressed the many incentive-based and structural barriers to the companies actually taking this approach, nor have we discussed the many difficult elements that would be involved in implementing it. This includes things such as how to successfully incorporate third party fact-checking and authoritative information sources, defining the appropriate role of AI or algorithmic content moderation tools (and implementing them responsibly and effectively), substantively specifying normative considerations, and scaling up the labor needed (with fair compensation and decent working conditions). Nevertheless, a systematic and normatively grounded approach can improve and elevate content moderation efforts by providing clearer ideas of what the goals are, how success should be defined and measured, and where ethical considerations should be taken into account.