Informational Quality Labeling on Social Media:
In Defense of a Social Epistemology Strategy

John P. Wihbey, Matthew Kopec & Ronald Sandler*

Abstract:

Social media platforms have been rapidly increasing the number of informational labels they are appending to user-generated content in order to indicate the disputed nature of messages or to provide context. The rise of this practice constitutes an important new chapter in social media governance, as companies are often choosing this new "middle way" between a laissez-faire approach and more drastic remedies such as removing or downranking content. Yet information labeling as a practice has, thus far, been mostly tactical, reactive, and without strategic underpinnings. In this paper, we argue against defining success as merely the curbing of misinformation spread. The key 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. The strategy we articulate emphasizes how the moderation system needs to improve 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 the platform — while also appropriately respecting sources, seekers, and subjects of information. A systematic and normatively grounded approach can improve content moderation efforts by providing clearer accounts of what the goals are, how success should be defined and measured, and where ethical considerations should be taken into consideration. We consider implications for the policies of social media companies, propose new potential metrics for success, and review research and innovation agendas in this regard.

***Authors:**

John P. Wihbey is an assistant professor of journalism and media innovation at Northeastern University where he heads the graduate programs in the School of Journalism. He is the author of *The Social Fact: News and Knowledge in a Networked World* (MIT Press, 2019). He is a faculty affiliate at the Northeastern Ethics Institute, the School of Public Policy and Urban Affairs, and the Center for Law, Innovation and Creativity (CLIC).

Matthew Kopec is the Associate Director of the Ethics Institute at Northeastern University. He has held teaching and/or research positions at The Australian National University, The Centre for Applied Philosophy and Public Ethics (Canberra, Australia), Northwestern University, and The University of Colorado - Boulder. His research covers topics at the intersections of social epistemology, applied ethics, and philosophy of science.

Ronald Sandler is a professor of philosophy, Chair of the Department of Philosophy and Religion, and Director of the Ethics Institute at Northeastern University. His primary areas of research are ethical theory, ethics and emerging technologies, and environmental ethics.

Informational Quality Labeling on Social Media:
In Defense of a Social Epistemology Strategy

John P. Wihbey, Matthew Kopec & Ronald Sandler [1]

[1] Acknowledgement: 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.

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. 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.[2] 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.[3] 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 policies—such as Facebook's "Community Standards" or "The Twitter

[2] 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.

[3] 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. *Quarantined Subreddits*, REDDIT, https://www.reddithelp.com/hc/en-us/articles/360043069012 (last visited Mar. 27, 2021).

Rules"—also embody this tactical conception of information quality labeling.[4] 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 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[5] 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.[6]

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 placed on how they currently work and the associated

---

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

[5] "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 (A Bradford Book, 2007); ALVIN GOLDMAN & DENNIS WHITCOMB, SOCIAL EPISTEMOLOGY: ESSENTIAL READINGS (Oxford University Press US, 2011); Alvin Goldman and Cailin O'Connor, *Social Epistemology*, THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta, ed., 2019), https://plato.stanford.edu/archives/fall2019/entries/epistemology-social/.

[6] Kay Mathiesen, *Informational Justice: A Conceptual Framework for Social Justice in Library and Information Services,* 64 Library Trends 198 (2015).

problems, issues, and challenges. In Section Two, we discuss why a systematic content labeling approach begins with articulating the values and goals of the 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.[7]

1. Section One: Defining the problem

A. Complex mechanics

Content moderation can be defined as the "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[8] 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 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.[9] 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.

---

[7] 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*.

[8] James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42 (2015).

[9] 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 U.S. of a new community-driven approach to help address misleading information"—that allow devolved moderation structures within a platform's larger centralized approach. *See* Keith Coleman, *Introducing Birdwatch, a community-based approach to misinformation*, TWITTER, (Jan. 24, 2021) https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.

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.[10] Hate speech, violent extremism, harassment, nudity, and self-harm are some of the many categories often subject to heavy moderation and removal. Some of this moderation is mandated by long-standing laws, such as those relating to copyright violations (e.g., the Digital Millennium Copyright Act, or DMCA),[11] while some newer laws globally, such as Germany's Network Enforcement Act, or NetzDG, are also increasingly mandating that social media companies remove defamatory content and hate speech.[12]

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 (e.g., 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.[13]

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 interest that is

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

[11] 17 U.S.C. § 512.

[12] Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law*, (Transatlantic High Level Working Group On Content Moderation Online And Freedom Of Expression Series, 2019).

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

> newsworthy or otherwise enables individual expression on topics of social importance.[14]

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 cursory human review. Edge cases are then escalated up to content review teams that have increasingly more policy oversight and authority.[15] 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 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.[16] After an initial refusal to release data about enforcement of community guidelines, companies such as YouTube, Twitter, Facebook/Instagram started reporting in 2018 more statistical information about their overall moderation efforts. These reports may include the total volume of content seeing moderation; the prevalence of categories such as hate speech on their platforms; and the degree of preemptive, algorithmic actions taken before violating content is widely shared.[17]

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.[18] Third-party fact-checkers 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.

---

[14] *Community Guidelines: Introduction*, TIKTOK (Dec. 2020), https://www.tiktok.com/community-guidelines?.

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

[16] Mark Scott & Laura Kayali, *What happened when humans stopped managing social media content*, POLITICO (Oct. 21, 2020, 5:56 PM), https://www.politico.eu/article/facebook-content-moderation-automation/.

[17] Daphne Keller & 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 Persily & Joshua A. Tucker eds., 2020). For an example of reporting, see *Community Standards Enforcement*, FACEBOOK, https://transparency.facebook.com/community-standards-enforcement (last visited Dec. 23, 2020).

[18] *Evaluating Platform Election-Related Speech Policies*, ELECTION INTEGRITY PARTNERSHIP (Oct. 28, 2020), https://www.eipartnership.net/policy-analysis.

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.[19] 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 2020 U.S. 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 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 goal of free and fair elections. The policy implementation and tactics employed vary widely.[20] 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.[21]

B. 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.[22] 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 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 the user-generated content they host.[23] 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

---

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

[20] *Evaluating Platform Election-Related Speech Policies*, *supra* note 18.

[21] Emily A. Vogels et al., *Most Americans Think Social Media Sites Censor Political Viewpoints*, PEW RESEARCH CENTER (Aug. 19, 2020) https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

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

[23] 47 U.S.C. § 230.

targeted for threats and abuse because of race or gender—and thereby constitutes a "discriminatory design" that disadvantages the most vulnerable in society.[24]

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.[25] Employees within Facebook, for example, have been reportedly split bitterly over how to balance the demands for growth with the need to maintain informational quality, civility, and safety on the platform.[26]

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.[27]

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

---

[24] Olivier Sylvain, *Discriminatory Designs on User Data*, *in* THE PERILOUS PUBLIC SQUARE: STRUCTURAL THREATS TO FREE EXPRESSION TODAY 181 (David E. Pozen ed., 2020); Danielle Keats Citron, *Section 230's Challenge to Civil Rights and Civil Liberties*, *in* THE PERILOUS PUBLIC SQUARE, *supra* at 200. 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*, THE BOSTON GLOBE (Dec. 18, 2020) https://www.bostonglobe.com/2020/12/18/opinion/social-media-broke-america-heres-how-fix-it/.

[25] Caplan, *supra* note 8.

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

[27] Garrett Morrow et al., *The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation* (Northeastern University Ethics Institute, Working Paper, Dec. 3, 2020).

moving away from the controlling ideas for news and information of the twentieth century embodied in Justice Holmes's 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."[28] 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.[29] 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—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. [30]This would include Section 230, enacted years 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 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. [31] 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 contemplate a need to require new forms of disclosure and context to mitigate the pathologies of a more wide-open system of communication.[32] Yet it has taken two decades for many to realize the extent to which the old paradigm has been overturned. Novel problems may now require a substantial rethinking of approaches and policy tools.

---

[28] *Abrams v. United States*, 250 U.S. 616, 624 (1919).

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

[30] For a discussion of the idea of policy drift more broadly, see J.S. Hacker, P. Pierson & K.A. Thelen, *Drift and Conversion: Hidden Faces Of Institutional Change*, *in* ADVANCES IN COMPARATIVE-HISTORICAL ANALYSIS 180 (J. Mahoney & K. A. Thelen eds., 2015).

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

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

C. Moderation and labeling challenges

Social media companies are now pouring millions, if not billions, of dollars into content moderation.[33] 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 to defend their right to exercise editorial control of their platform content, although these may be asserted on questionable grounds.[34]

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 nimbler and "softer" approaches such as warning labels, knowledge panels, source transparency buttons, and other "metadata" instruments, or information about information. [35] 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, the concept is immature in other domains such as consumer products, food, pharmaceuticals, and even media- and information-driven spaces such as the entertainment industry.

There exists a major body of research literature relating to information labeling and disclosure in the context of public regulation and governance,[36] but few have studied how such insights might be operationalized in a social media context. Facebook announced in 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 scholarly works[37] 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—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.[38]

---

[33] Janko Roettgers, *Mark Zuckerberg Says Facebook Will Spend More Than $3.7 Billion on Safety, Security in 2019*, NASDAQ (Feb. 5, 2019, 12:32 PM) https://www.nasdaq.com/articles/mark-zuckerberg-says-facebook-will-spend-more-37-billion-safety-security-2019-2019-02-05.

[34] Kyle Langvardt, *Platform Speech Governance and the First Amendment: A User-Centered Approach* (The Digital Social Contract: A Lawfare Paper Series, 2020).

[35] Morrow et al., *supra* note 27.

[36] CASS R. SUNSTEIN, TOO MUCH INFORMATION: UNDERSTANDING WHAT YOU DON'T WANT TO KNOW (2020).

[37] 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).

[38] Morrow et al., *supra* note 27.

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 the companies' 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."[39] Some critics claim that labels are more a form of public relations, and less a substantive attempt to deal with the problem of misinformation.[40]

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.[41] On questions specific to the labeling of inaccurate information, there are deep partisan divisions. Conservative-leaning respondents were overwhelmingly likely to doubt the legitimacy and intentions of social media labeling efforts while liberal respondents were split in terms of confidence in the companies to make these decisions.[42] 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.[43]

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 such as outsourced work of other firms who help with the moderation tasks set up by the companies. 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, and accuracy.[44] The consequences of the more heavy-handed content moderation decisions such as takedowns and removal have seen some

[39] Mike Ananny & Tarleton Gillespie, *Public Platforms: Beyond the Cycle of Shocks and Exceptions*, (The Internet, Policy & Politics Conferences, University of Oxford, 2016).

[40] 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/.

[41] Vogels et al., *supra* note 21.

[42] *Id*.

[43] Sarah Myers West, *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms*, 20 New Media & Soc'y 4366 (2018).

[44] Tarleton Gillespie, *Content moderation, AI, and the question of scale*, 7 Big Data & Soc'y (2020).

amount of careful study, although public understanding remains limited because of a lack of full transparency about the platforms' work in this respect.[45]

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.[46] 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.[47]

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 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 have shown how these systems sometimes result in arbitrary decisions with little remedy.[48] Content moderation teams may need to be significantly expanded and the work function raised to a higher-status role within companies.[49]

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

---

[45] 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 Persily & Joshua A. Tucker eds., 2020).

[46] 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 (Mar. 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.

[47] 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.

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

[49] PAUL M. BARRETT, WHO MODERATES THE SOCIAL MEDIA GIANTS?, (Center For Business At New York University, June 2020).

potential competitors and locking in the advantages of the existing mega-platforms.[50]

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 seemingly piecemeal individual instances of misinformation: downranking some posts, removing others, and labeling or "fact-checking" still others. This 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.[51]

There is a 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."[52] Yet if misinformation is not labeled by these platforms according to a transparent and consistently enforced policy, surely the public will not be much 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?

---

[50] Tarleton Gillespie et al., *Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates*, 9 Internet Pol'y Rev. 1 (2020).

[51] 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/.

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

2. 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.[53] 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).[54] 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;[55] and Facebook the techno-social system is designed in a way that resists evaluating informational quality. Their professed ideal is neutrality.

---

[53] 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." *Resources: FAQ's*, FACEBOOK INVESTOR RELATIONS (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." *Contact: FAQ's*, TWITTER INVESTOR RELATIONS (2021) https://investor.twitterinc.com/contact/faq/.

[54] *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/.

[55] 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.

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.[56] 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 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. Companies cannot claim neutrality with respect to informational quality and 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.[57] But in this context it is worth emphasizing

---

[56] Kay Mathiesen & Don Fallis, *Information Ethics and the Library Profession*, *in* HANDBOOK OF INFORMATION AND COMPUTER ETHICS 221(Kenneth Einar Himma & Herman T. Tavani eds., 2008).

[57] ARNOLD PACEY, THE CULTURE OF TECHNOLOGY (1985); Langdon Winner, *Do Artifacts Have Politics?*, 109 Daedalus 121 (1980); Langdon Winner, *Technologies as Forms of Life*, *in* EPISTEMOLOGY, METHODOLOGY, AND THE SOCIAL SCIENCES (Robert S. Cohen & Marx W. Wartofsky eds., 2013).

that this applies in particular to values related to quality and 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 and recent events, such as increasing connectivity while avoiding harms to individuals (these are the ones 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.

3. The need for a social epistemic approach

A. 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).[58] Any proper content moderation strategy

[58] Regina Rini, *Fake News and Partisan Epistemology*, 27 Kennedy Inst. Ethics J. E–43 (2017); C. Thi Nguyen, *Cognitive Islands and Runaway Echo Chambers: Problems for Epistemic Dependence on Experts*, 197 Synthese 2803 (2020); C. Thi Nguyen, *Echo Chambers and*

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 or 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,[59] the study of norms of human thought, rationality, justification, and knowledge prior to the 1980s tended to focus on the sole inquirer, attempting to 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 have been taught to reason by our mentors or society, and our informational positions and relationships.[60] 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.[61]

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 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,[62] this labeling strategy wasn't

Epistemic Bubbles, 17 Episteme 141 (June 2020); Don Fallis & Kay Mathiesen, *Fake News Is Counterfeit News*, 0 Inquiry 1 (2019).

[59] Here, we have in mind the likes of C.S. Peirce, Émile Durkheim, Thomas Kuhn, and Paul Feyerabend, among others.

[60] THE ROUTLEDGE HANDBOOK OF SOCIAL EPISTEMOLOGY (Miranda Fricker et al. eds., 2019).

[61] For some approaches in epistemology that focus explicitly on improving the information environment see Alvin Goldman, *Systems-oriented social epistemology*, 3 Oxford Studies in Epistemology 189 (2010); Don Fallis, *On Verifying the Accuracy of Information: Philosophical Perspectives*, 52 Library Trends 463 (2004); Shane Ryan, *Epistemic Environmentalism*, 43 J. Philosophical Rsch. 97 (2018).

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

able to contain the spread of bogus cures (like Vitamin C),[63] conspiracy theories concerning the origin of the virus (like the 5G conspiracy theory),[64] or anti-vaccination information.[65] 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., 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.[66] 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.[67] 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.[68] Users are often presented with information from an increasingly homogenous set of viewpoints.[69]

---

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

[64] Wasim Ahmed et al., *COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data*, 22 J. Med. Internet Rsch. e19458 (2020); 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.

[65] Talha Burki, *The Online Anti-Vaccine Movement in the Age of COVID-19*, 2 The Lancet Digital Health e504–5 (Oct. 1, 2020).

[66] On the epistemology of expertise, see SANFORD C. GOLDBERG, RELYING ON OTHERS: AN ESSAY IN EPISTEMOLOGY (Oxford University Press, 2010); Alvin I. Goldman, *Experts: Which Ones Should You Trust?*, 63 Phil. & Phenomenological Rsch. 85 (2001); C. Thi Nguyen, *Cognitive Islands and Runaway Echo Chambers*, *supra* note 59; James Owen Weatherall & Cailin O'Connor, *Endogenous Epistemic Factionalization*, Synthese (2020) (on file with Dep't of Logic and Phil. of Sci. at U.C. Irvine).

[67] On the epistemology of testimony see Miranda Fricker, *Group Testimony? The Making of a Collective Good Informant*, 84 Phil. & Phenomenological Rsch. 249 (2012); DEBORAH TOLLEFSEN, GROUPS AS AGENTS (John Wiley & Sons eds., 2015); JENNIFER LACKEY, LEARNING FROM WORDS: TESTIMONY AS A SOURCE OF KNOWLEDGE, (Oxford University Press, 2008); Karen Frost-Arnold, *Trustworthiness and Truth: The Epistemic Pitfalls of Internet Accountability*, 11 Episteme 63 (2014).

[68] On the epistemic norms of assertion, see SANFORD GOLDBERG, ASSERTION: ON THE PHILOSOPHICAL SIGNIFICANCE OF ASSERTORIC SPEECH, (Oxford University Press, 2015); SANFORD C. GOLDBERG, TO THE BEST OF OUR KNOWLEDGE: SOCIAL EXPECTATIONS AND EPISTEMIC NORMATIVITY (Oxford University Press, 2018); John Turri, *Truth, Fallibility, and Justification: New Studies in the Norms of Assertion*, Synthese 1 (2020); JESSICA BROWN & HERMAN CAPPELEN, ASSERTION: NEW PHILOSOPHICAL ESSAYS, (Oxford University Press, 2011).

[69] On the epistemic concerns raised by homogenous evidence sources, see Kenneth Boyd, *Epistemically Pernicious Groups and the Groupstrapping Problem*, 33 Soc. Epistemology 61

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.[70]

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.[71] This is evident in some Trump supporters' beliefs that Trump 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).[72] More recently, network epistemologists have been working on ways to modify social networks in order to increase the likelihood of obtaining certain epistemic goals.[73] And as Neil Levy and Mark Alfano have 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.[74] 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

(2019); Engin Bozdag & Jeroen van den Hoven, *Breaking the Filter Bubble: Democracy and Design*, 17 Ethics & Info. Tech. 249 (2015).

[70] On the epistemology of filter bubbles and echo chambers, see Nguyen, *Echo Chambers and Epistemic Bubbles, supra* note 59; C Thi Nguyen, *Why It's as Hard to Escape an Echo Chamber as It Is to Flee a Cult*, AEON (Apr. 9, 2018), https://aeon.co/essays/why-its-as-hard-to-escape-an-echo-chamber-as-it-is-to-flee-a-cult; Benjamin Elzinga, *Echo Chambers and Audio Signal Processing*, Episteme 1 (2020); CAILIN O'CONNOR AND JAMES OWEN WEATHERALL, THE MISINFORMATION AGE: HOW FALSE BELIEFS SPREAD (2019).

[71] This kind of scenario is commonly discussed in the literature on the "false consensus effect." *See* M. Wojcieszak, *False consensus goes online: Impact of ideologically homogeneous groups on false consensus*, 72 Pub. Op. Q. 781 (2008).

[72] 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. For Info. Sci. & Tech. 1662 (2008); Alexander Bird, *Social Knowing: The Social Sense of 'Scientific Knowledge*, 24 Phil. Persps. 23 (2010); Søren Harnow Klausen, *Group Knowledge: A Real-World Approach*, 192 Synthese 813 (2014).

[73] Conor Mayo-Wilson et al., *Wisdom of Crowds versus Groupthink: Learning in Groups and in Isolation*, 42 Int'l J. Game Theory 695 (2013); Kevin J. S. Zollman, *The Epistemic Benefit of Transient Diversity*, 72 Erkenntnis 17 (2009).

[74] Neil Levy & Mark Alfano, *Knowledge From Vice: Deeply Social Epistemology*, 129 Mind 887 (2020).

clear about whether they are actually endorsing what they share.[75] 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.

B. 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. 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."[76] Here one can see that Twitter is suggesting that the labels were efficacious in reducing the spread of the false claims.[77]

Even assuming this is true, that the labels significantly reduced retweeting and so reduced the spread of the president's misinformation, Twitter's rationale and approach nevertheless amount to what we have referred to as a very individualistic 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 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

[75] Rini, *supra* note 59; Rachel Sterken et al., *On Retweeting* (2019) (Manuscript, forthcoming).

[76] *An Update on Our Work around the 2020 US Elections*, TWITTER (Nov. 12, 2020) https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html.

[77] Paul Mena, *Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook,* 12 Pol'y & Internet 165 (2020; 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/.

challenge from a social epistemology perspective is not "how to make truth travel faster than lies,"[78] 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 in the wake of the January 6, 2021 attack on the U.S. Capitol on the basis of inciting violence.[79] 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.[80] The still 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 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 labeling 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 this has 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.[81]

---

[78] Geoffrey A. Fowler, *supra* note 78.

[79] Facebook Newsroom (@fbnewsroom), TWITTER (Jan. 7, 2021, 11:01 AM) https://twitter.com/fbnewsroom/status/1347211647245578241.

[80] Jeff Horwitz & Keach Hagwy, *Parler Makes Play for Conservatives Mad at Facebook, Twitter*, THE WALL STREET JOURNAL (Nov. 15, 2020, 1:05 AM) https://www.wsj.com/articles/parler-backed-by-mercer-family-makes-play-for-conservatives-mad-at-facebook-twitter-11605382430.

[81] 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

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),[82] then that is not success.[83] It is social epistemic success that is needed, and that might mean more robust labels with links to correct sources are preferable to suppression.[84] 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 personal and social costs associated with that.[85]

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.[86]

---

unrelenting personalization of content that erodes shared knowledge, the absence of a consistent, intelligible and research-based labeling strategy, and so on.

[82] Cass R. Sunstein & Adrian Vermeule, *Conspiracy Theories: Causes and Cures*, 17 J. Pol. Phil. 202 (2009).

[83] Stephan Lewandowsky et al., *Misinformation and Its Correction: Continued Influence and Successful Debiasing*, 13 Psych. Sci. 106 (2012).

[84] Morrow et al., *supra* note 27, 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 & David Lazer, *Public health and online misinformation: challenges and recommendations*, 41 Ann. Rev. Pub. Health 433 (2020); Briony Swire-Thompson et al., *Searching for the backfire effect: Measurement and design considerations*, 9 J. Appl. Res. Mem. Cogn. 286 (2020).

[85] It is 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 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.

[86] 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*, NIEMANLAB (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/.

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 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.[87] 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,[88] 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 misinformation.[89] But our hope is that the preceding discussion elucidates what it means to approach informational quality content moderation from a social

[87] 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. As is now well accepted, however, 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 et al., *The Weirdest People in the World?*, 33 Behav. & Brain Sci. 61 (2010).

[88] Morrow et al., *supra* note 27.

[89] Rini, *supra* note 59; Nguyen, *supra* note 59; Fallis and Mathiesen, *supra* note 59; Sterken et al., *supra* note 76.

epistemology perspective, and demonstrates how it provides a useful perspective for analyzing the problem of content moderation and developing and evaluating candidate approaches to addressing it.

4. 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 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.[90]

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

[90] Ifeoma Ajunwa, *The paradox of automation as anti-bias intervention*, 41 Cardozo L. Rev. 54 (2020); Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, 3 Big Data & Soc'y 1 (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.

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 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 whether or not they themselves are purveyors of misinformation by reducing their ability to share information, 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,[91] 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 so be problematic.[92] 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.[93]

---

[91] Daniel Susser, *Predictive Policing and the Ethics of Preemption*, *in* THE ETHICS OF POLICING: AN INTERDISCIPLINARY PERSPECTIVE (Ben Jones & Eduardo Mendieta eds., NYU Press) (forthcoming).

[92] 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* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017, 5 AM) https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms; SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Mona Sloane & Emanuel Moss, *AI's social sciences deficit*, 1 Nature Machine Intelligence 330 (2019); RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (2019); MEREDITH BROUSSARD, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (2018).

[93] For in depth treatments of related concerns over epistemic distributive justice see Faik Kurtulmus and Gurol Irzik, *Justice in the Distribution of Knowledge*, 14 Episteme 129 (2017);

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., individuals that posts or claims are about);[94] 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.[95] 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 labeling 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 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

---

Don Fallis, *Epistemic Value Theory and the Digital Divide*, *in* INFORMATION TECHNOLOGY AND SOCIAL JUSTICE 29 (E. Rooksby & J. Weckert eds., 2007).

[94] Johannes J. Britz, *Making the Global Information Society Good: A Social Justice Perspective on the Ethical Dimensions of the Global Information Society*, 59 J. Am. Soc'y for Info. Sci. & Tech. 1171 (2008); Kay Mathiesen, *Access to Information as a Human Right* (2008); Kay Mathiesen, *Informational Justice: A Conceptual Framework for Social Justice in Library and Information Services*, 64 Library Trends 198 (2015); Kay Mathiesen, *The Human Right to Internet Access: A Philosophical Defense*, 18 The Int'l Rev. Info. Ethics 9 (2012).

[95] For example, imagine how different the world would be, from a social justice perspective, if content moderation regimes had inadvertently suppressed the Arab Spring.

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).[96] 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 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,[97] 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 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 limited in 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

---

[96] MIRANDA FRICKER, EPISTEMIC INJUSTICE: POWER AND THE ETHICS OF KNOWING (1st ed. 2009); Rachel McKinnon, *Epistemic Injustice*, 11 Phil. Compass 437 (2016).

[97] 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.

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.

5. 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 is 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 and 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 and 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. It should therefore 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 modify 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 fact-checkers improve outcomes.[98] For example, it would be very useful to conduct experiments on platforms that vary in 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 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),[99] as well as norm engineering around retweeting.[100] 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 for them) 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.

---

[98] 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 et al., *Tackling misinformation: What researchers could do with social media data*, 1 Harv. Kennedy Sch. Misinformation Rev. (2020).

[99] Rini, *supra* note 59.

[100] Sterken et al., *supra* note 76.

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 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.