Harvard Kennedy School Misinformation Review[1]
January 2020, Volume 1, Issue 1

DOI: https://doi.org/10.37016/mr-2020-005
Website: misinforeview.hks.harvard.edu

Commentary

Answering impossible questions: Content governance in an age of disinformation

The governance of online platforms has unfolded across three eras – the era of Rights (which stretched from the early 1990s to about 2010), the era of Public Health (from 2010 through the present), and the era of Process (of which we are now seeing the first stirrings). Rights-era conversations and initiatives amongst regulators and the public at large centered dominantly on protecting nascent spaces for online discourse against external coercion. The values and doctrine developed in the Rights era have been vigorously contested in the Public Health era, during which regulators and advocates have focused (with minimal success) on establishing accountability for concrete harms arising from online content, even where addressing those harms would mean limiting speech. In the era of Process, platforms, regulators, and users must transcend this stalemate between competing values frameworks, not necessarily by uprooting Rights-era cornerstones like CDA 230, but rather by working towards platform governance processes capable of building broad consensus around how policy decisions are made and implemented. Some first steps in this direction, preliminarily explored here, might include making platforms "content fiduciaries," delegating certain key policymaking decisions to entities outside of the platforms themselves, and systematically archiving data and metadata about disinformation detected and addressed by platforms.

Authors: John Bowers (1), Jonathan Zittrain (1)
Affiliations: (1) Harvard Law School, USA
How to cite: Bowers, J., & Zittrain, J. (2020). Answering impossible questions: Content governance in age of disinformation. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(1).
Received: November 13th, 2019. Accepted: December 24th, 2019. Published: January 14th, 2020.

Today's dominant internet platforms are built on the notion that people with no audience can, through a snowball of engagement, quickly garner one if they have something compelling to say. The new monetary, political, and reputational opportunities made possible by this ethos have given rise to an art and science, supercharged by targeted advertising, around optimizing the placement and spread of content for fun and profit. In the course of ushering in a new era of e-commerce, creative production, and online discourse, this "craft" of virality has placed enormously powerful new tools in the hands of propagandists seeking to manipulate online audiences through disinformation. From election meddling and political polarization to mob violence and genocide, headline-grabbing real-world harms arising from abuse of platforms' affordances have led lawmakers and advocates to call for accountability and change.

In the U.S., much of the conversation around what platforms can be held accountable for has centered on limitations imposed by a legal framework implemented in 1996 and, since then, continuously upheld as a keystone of the digital content governance environment: the modestly-named "section 230" of the

[1] *A publication of the Shorenstein Center for Media, Politics and Public Policy, at Harvard University, John F. Kennedy School of Government.*

Communications Decency Act. CDA 230 insulates online intermediaries like today's platforms from liability relating to most user content and behavior that doesn't run afoul of criminal law, including much of what we would today label as disinformation (47 U.S.C. § 230). Its protections for online speech, designed to safeguard then-nascent spaces against external coercion or interference, apply even in cases where intermediaries are actively involved in moderating user content – indeed, that's where CDA 230 was meant to most affect then-prevailing law (47 U.S.C. § 230). By providing such a broad legal shield and embedding it in a federal statute that restrained what individual states could do, CDA 230 placed the dynamics of platform liability in amber for the sake of protecting forums for online speech. What might otherwise have been a decades-long process of common law development aimed at defining the specific contours of platform liability with respect to harmful content was instead determined in short order. CDA 230 is the signature product of what we might call the "Rights" era of internet governance, a period – stretching from the dawn of the commercial internet of the early 1990s to about 2010 – during which public and regulatory conversations focused almost exclusively on protecting a maturing sphere of internet discourse from external coercion, whether corporate or governmental (Zittrain, 2019).

As this Rights-era intermediary liability framework has remained fixed in place, the public's sense of what platforms ought to be held accountable for has shifted tectonically (Zittrain, 2019). Contemporary notions of platform responsibility reflect an awareness of population-level harms foreign to the world of 1996, like those enabled by weaponized virality. They also recognize the jaw-dropping personalization capabilities of today's online platforms, maintained by companies that are approaching trillion-dollar valuations based on their ability to sort, filter, and target content. These sensibilities have, since about 2010, ushered in a new era: that of Public Health. In this era, the speech protection interests of the Rights era are competing in public discourse against concerns relating to concrete and measurable aggregate harms to individuals and institutions – and, in particular, the corrosive social and political effects of disinformation (Zittrain, 2019). This new calculus has found little footing in still-dominant Rights-era doctrine, leading numerous policymakers and commentators to brand CDA 230 as a dangerous, outdated subsidy to some of the world's richest and most powerful companies (Laslo, 2019).[2] But it seems doubtful that even a full repeal of CDA 230 would facilitate, much less impose, such accountability. Some experts like Stanford's Daphne Keller have argued that such a repeal would also endanger the still-relevant concerns of the Rights era, making the hosting of online speech a risky proposition, and likely restricting many forms of unobjectionable – and Constitutionally-protected– speech (Zittrain, 2019; Keller, 2018).

Beyond the clash between rights and public health, a third era is in demand, one of Process, which should focus on the development of broadly-legitimated mechanisms for achieving workable compromises among rights, public health, and any number of other emerging considerations. These mechanisms will need to break from the opaque and inwards-looking content governance models of today. By focusing on building credibility around the means by which content governance decisions are reached, the era of Process will lean into the ambiguities and tradeoffs involved in balancing rights and public health interests, enabling decisive decision-making even in the face of unavoidable substantive controversy.

From rights to public health

It is easy to champion free speech when the stakes of harmful speech are small and largely hypothetical

[2] Criticism of CDA 230 has, by and large, come in two distinct strains. Some policymakers – particularly Republicans like Senator Josh Hawley (R-Mo) – have argued that CDA 230 enables politically biased content moderation on the part of platforms. The criticism that characterizes the Public Health argument, however, takes the view that CDA 230 shields platforms that fail to moderate extensively enough, removing incentives to police markedly harmful content.

– harms just enough to elicit a solemn acknowledgement that freedoms sometimes come with a cost. When harms arising from speech become serious and commonplace, however, such permissive standards press for much greater justification. The decades following the passage of CDA 230 saw an escalation of just this kind, powered by a dramatic shift in how online content found its audience. In opening the door to viral disinformation and other speech-related ills, this shift laid the groundwork for the dawn of the Public Health era of internet governance.

In the late 1990s and early 2000s, the world in which CDA 230 was passed, online content was typically divided into two relatively neat categories, reflecting what were experienced as two distinct channels of content provision online (Zittrain, 2019). On one hand, there was content developed and disseminated by traditional publishers that had moved into the online space. On the other, there was user-generated content, with the ungainly shorthand of "UGC": the contributions of individuals, sometimes anonymous, to message boards, blogs, comment sections, and personal website platforms like Geocities. CDA 230 was designed to inoculate platforms which hosted this second category of content, whether as their sole purpose or otherwise, from liability arising from it.[3] UGC was, at the time, largely identifiable from where it was placed and how it looked: a string of comments, say, at the end of a more magisterial article. The supposed clarity of this demarcation lessened the perceived stakes of harms arising from UGC, since it occupied its own less-serious lane – browsing a comments section, conspiracy blog, or political forum was clearly separate from reading headlines on the website of an established publication.[4]

And then, with the rise of a handful of dominant internet platforms that indexed and pointed to everything else online, the line between UGC and content from traditional publishers blurred. The platforms introduced a new model of content consumption built around feeds of material, like Google Search results or a Facebook Newsfeed. The feed contents were ranked by algorithms, optimizing, at least initially, for user engagement above all else, and making little visible distinction between different forms of content. They quickly flattened the landscape of internet content, presenting even the most rough-hewn of UGC in a common stream and on par with meticulously fact-checked articles by professional journalists. Under this consolidated circulation model, anything can rapidly go viral with enough engagement, regardless of bona fides or provenance.[5]

The influence of UGC can now readily match or exceed that of content produced by professionalized newsrooms, but, thanks to CDA 230, platforms hosting harmful UGC remain every bit as inoculated from liability for its effects as they were two decades ago.

Disinformation campaigns blend right into this mélange. The move towards source-agnosticism opened the door for propagandists' weaponization of platform-based content dissemination. As Facebook CEO Mark Zuckerberg himself pointed out in a 2018 blog post, users have a tendency to engage with and amplify content engineered to sensationalize, defame, or mislead (Zuckerberg, 2018). Such users can readily be reached and influenced through the content distribution channels, organic or ad-driven, furnished by platforms. A now-familiar litany of disinformation-related harms is worth restating: Russian influence campaigns targeting the 2016 US election, however, disputed their actual impact might be,

---

[3] *The New York Times*, for example, can be held liable for defamatory statements included in articles it self-publishes on its website at nytimes.com, but cannot be held liable for defamatory statements made by others through its comments sections, or even published as letters to the editor. (Material in the paper edition, however, is not shielded.)

[4] This assumption is reflected in a substantial body of literature from the early to mid-2000s, which engages with questions relating to "participatory journalism."

[5] In his book *Amusing Ourselves to Death*, written in 1985 about the power and culture of television, Neil Postman lamented the rise of a world in which broadcast news anchors would shift from one story to another with a generic segue intoning, "And now … this." Postman decried it as "a world of fragments, where events stand alone, stripped of any connection to the past, or to the future, or to other events – that all assumptions of coherence have vanished" (110). His fears have come to their logical conclusion in an online content ecosystem that intentionally intermingles everything, from highbrow feature stories and cat videos to distant friends' musings on tax reform or the death of a parent.

stand to undermine our faith in online discourse.[6] Vaccine misinformation has contributed to the re-emergence of previously eradicated diseases (Robeznieks, 2019). Social media content intended to radicalize has fueled lethal violence in India and Myanmar, among other places. In 2017, the St. Petersburg-based Internet Research Agency even attempted, by way of Facebook's Events feature, to convene a town hall meeting in Twin Falls, Idaho to address what it called, in the guise of locals, a "huge upsurge of violence towards American citizens" carried out by refugees (Shane, 2017).

However lurid these examples might be, disinformation-related harms often arise less from, say, how many people went to the spurious Twin Falls town hall meeting, and more from the accumulation of chronic patterns of low-level but nonetheless toxic user behavior, coordinated or otherwise. For example, while a given tweet skeptical of vaccination might not seem particularly concerning in isolation, it can become so as part of a coordinated influence campaign aimed at reducing vaccination rates. Public recognition of the real-world costs of this slow poisoning has brought regulators and others to speak quite plainly in terms of what we might call a "Public Health" model of content governance, which concerns itself with modelling and addressing these aggregate effects.[7] Rather than simply defining through code, architecture, and policy what users can and cannot post, comment, or share, platforms are now being asked to don the epidemiologist's hat and mitigate specific and contextual harms to norms and institutions arising from interactions between users on a massive scale.

Process as transcendence

Embarking on a constructive project to make content governance on the platforms accountable for coherently balancing rights and public health interests will require a turn towards a third era: Process. The inwards-looking, largely public relations-oriented content governance models so widely deployed today are unsatisfying. That's in part because they don't seem to be working very well, but also because they assume that sensitive interests of individuals and society – like, say, the preservation of democratic norms in the face of disinformation – can be given reasonable treatment under corporate "customer service" processes architected to defuse PR pressure and protect profitability. As a reaction to mounting public discomfort with this incongruity, the platforms have begun to grasp at new means of "sectioning off" some decisions that implicate sensitive interests for special treatment. We have seen a crush of newly appointed Chief Ethics Officers and ethics review boards, all intended to represent user interests in nominally non-financial terms.[8]

These efforts are, by and large, weak medicine. Responsibility for addressing content-related harms like disinformation, in all of their significance and complexity, cannot be sidled onto an external advisory board without binding power to shape firm behavior. Ethics officers or internal review committee,

---

[6] Indeed, the 2019 Edelman Trust Barometer shows that only 34% of Americans trust social media as a news source, and that 73% of respondents "worry about false information or fake news being used as a weapon."

[7] While the last couple of years has seen an unprecedented flurry of high-profile government-level investigations and hearings related to disinformation, content governance, and CDA 230's liability shield (some frighteningly uninformed), this shift towards a public health-motivated model of platform responsibility didn't happen overnight. In their book *The Offensive Internet: Speech, Privacy, and Reputation*, published in 2010, editors Saul Levmore and Martha Nussbaum present more than a dozen essays by major internet scholars, each reflecting on the abuses and governance missteps enabled by CDA 230's permissive intermediary liability regime. While the introduction to the book, written by Levmore and Nussbaum, unambiguously encourages readers to consider a CDA 230 rollback as a legally feasible step forward for the internet, the essays that follow offer an interesting blend of rights and public health considerations. And even as many of the essays call upon us to reconsider CDA 230's protections in the face of concrete and imminent harms unfolding on the population level, they maintain a strong, even primary focus on protecting speech rights.

[8] For example, Salesforce has appointed a "Chief Ethical and Humane Use Officer" to focus on the ethical implications of technological innovation. Microsoft has launched an AI and Ethics in Engineering and Research (AETHER) Committee aimed at supporting internal policymaking. Google subsidiary DeepMind is building out ethics teams, and subsidizing university research.

embedded as they are within firms' existing incentive and authority structures, provide a poor alternative. The principled balancing of rights and public health interests in content governance should be an end in itself, to be respected and pursued even at the cost of the shareholder. A key question prompting the era of Process, then, is whether accountable orientation can be achieved within the structures of the platforms as they currently exist. If not, responsibility for key aspects of content governance must take place at least in part outside of the platforms, at an organizational remove from their business interests.

The era of Process must consider both solutions that assign platforms new duties with respect to content governance, and those that would delegate important aspects of the content governance process outside of the platforms themselves. Innovations of both kinds are conceivable even without requiring fundamental revisions to intermediary liability law. By finding models for accountability capable of functioning in the shadow of CDA 230, we may be able to avoid the messiness – and potential jeopardization of rights and competition interests – implicated by a revision or repeal. The objective of these new models should be, first and foremost, to develop and build legitimacy around new ways of working through ambiguous and controversial content governance questions, particularly those having to do with tradeoffs between rights and public health interests. Building legitimacy is distinct from arriving at the substantively "right" answer; legitimacy is most needed when there simply will not be broad public consensus about what the right answer is.

Platforms as content fiduciaries

In exploring solutions internal to companies, under which platforms might retain their grasp on content governance while accepting toothier forms of accountability, the notion of a fiduciary duty between platforms and their users may be useful. The deployment of fiduciary doctrine to address the problems of big tech is currently being explored under the "information fiduciaries" proposal as a means of protecting user privacy,[9] and its core ideas may be similarly applicable to content governance (Balkin, 2016; Balkin & Zittrain, 2016). Platforms have access to vast stores of data about user preferences and tendencies, which can be operationalized to shape user behavior in ways that drive revenue. Users, on the other hand, usually lack detailed insight into the mechanics of the platforms upon which they rely. In other professional relationships where such knowledge asymmetries exist, as between doctor and patient, or lawyer and client, the law recognizes and accounts for the vulnerability of the disadvantaged party by assigning a fiduciary duty to the advantaged party. This duty holds that the fiduciary must not jeopardize the interests of her charge, even when that means subordinating her own interests. When a fiduciary fails to abide by this rule, she can face consequences.

Embracing a framework of platform accountability built on fiduciary relationships – whether adopted voluntarily on the part of platforms, or imposed through regulatory action – might offer a path towards accountable content governance compatible with the existing decision-making structures of the platforms. By establishing a duty to protect and uphold the interests of the user, a fiduciary approach would set standards for content governance in terms of the specific dynamics of interest and power in relationships between platforms as content curators and their users as content consumers. It would add a new, legally-enforceable sense of professional responsibility to platforms' handling of content-related harms.

A fiduciary model of content governance could force platforms to grapple with the specificities and nuances of such harms. Take, for example, the importance of user intentionality in crafting platform policy around disinformation. It would probably be reasonable under a fiduciary framework for a narrowly

[9] See also S.3744 ("Data Care Act of 2018").

crafted search engine query for "trollfarm.ru article proving pizzagate" to yield the (propagandistic) result the user was clearly aiming to surface, perhaps alongside countervailing results like fact-checks. But when a user turns to their social media feed out of boredom, or searching on more generic terms, surfacing that same piece of content from among a large inventory (because the user is likely to click on it when presented with it unasked) could mean exposing a potentially unwitting individual to manipulation. By analogy, when a library patron asks for Mein Kampf at the circulation desk, we expect, in large part as a matter of expressive freedom, that librarian will return with that hateful text in hand or direct the patron to a place where it can be found. At the same time, we would likely hope that an open-ended request for "something interesting to read" might turn up a different book.

Delegating content governance

Whether or not platforms responsible for carrying out content governance take on a fiduciary duty with respect to their users,[10] placing certain sensitive decisions in the hands of an external entity may prove an essential counterpart. Delegating policymaking processes around certain key questions to entities outside of the platforms themselves puts a measure of distance between judgement and implementation valuable both to the platforms and to the public. What platforms delegate, though, is only half of the equation – the era of Process will also require us to envision new entities capable of taking up those questions. Facebook's slowly-materializing Independent Oversight Board represents the beginnings of one such model. A charter for the board, released in September of 2019, suggests that it will be responsible for reviewing difficult content decisions escalated by users or by Facebook itself, and that its decisions will have precedential value. While Facebook's fledgling investment in this model represents an important step towards meaningful externalization, whether or not it will serve as an effective externalization model remains to be seen. And there are good reasons to be skeptical of whether the Board's declarations will be folded back into platform policymaking in an impactful and expedient way.

In attaining real legitimacy, the proof will be in the pudding. But to know whether these new delegation models are worthy of our trust, we will need new ways of making transparent the form and effects of changes in platform policy. In the case of disinformation, platforms might better enable this transparency by archiving information about disinformation they've detected, along with records of their efforts to track and take action against those pieces of content. Given the sensitivity of this data, it may make sense for platforms to delay its availability for a period of months, and restrict access to accredited – but fully independent – academic researchers.

Private-sector transparency efforts around disinformation have been plagued by the specific challenges of releasing data to researchers or the public in real time – such releases often risk resurfacing suppressed campaigns, or even providing adversaries with tactical insights. A long-term archival approach to disinformation data sharing would be no substitute for up-to-the-minute telemetry, but could prove much more immediately tractable, and would provide enormous value to researchers and platforms trying to understand the long -term evolution of both platform policy and disinformation campaigns themselves. Importantly, such an approach would enable the evaluation of platforms' process-oriented innovations – like adherence to a fiduciary model, or the delegation of decisionmaking – in terms of their concrete implementation. Regardless of how well-designed, duty-bound, or appropriately delegated a platform's disinformation policymaking might be, its outputs will never be truly validated in the public eye without thorough inspection.

[10] Indeed, moving key decision-making processes outside of the firm may be useful to a newly minted fiduciary struggling to develop coherent models of user interest. An external entity might be well-positioned to weigh in as to what would and wouldn't violate fiduciary duty.

Conclusion

Debates around the proper scope of platform immunities under provisions like CDA 230 will, in part, shape the strategies of regulators and platforms in navigating the murky waters of content governance. However, we cannot let them keep us from thinking beyond the bounds of laws already written and structures already built, no matter how essential debates around the oft-ambivalent fruits of the Rights era and the new concerns of the Public Health era might be. We are on the brink of a Process era of internet governance. Moving away from a frustrating and unproductive clash between those two seemingly incompatible discourses, Rights and Public Health, will mean, to start, adopting governance models which embody a new "professionalism" in content governance. Whether it takes shape through a reorientation of platforms' content governance decision-making processes defined by user interests; through the delegation of certain decisions outside of the platforms themselves; or, perhaps ideally, through some combination of both, this professionalism, predicated on new forms of platform accountability, represents our best chance of building legitimacy around how content is sorted, filtered, and ranked.

Bibliography

Balkin, J. M. (2016). Information fiduciaries and the first amendment. *UC Davis Law Review, 49*(4), 1183-1234. https://lawreview.law.ucdavis.edu/issues/49/4/Lecture/49-4_Balkin.pdf

Balkin, J. M., & Zittrain, J. (2016, October 3). A grand bargain to make tech companies trustworthy. *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/

Communications Decency Act of 1996, 47 U.S.C. § 230 (1996). https://www.law.cornell.edu/uscode/text/47/230

Keller, D. (2018, April 6). *Toward a clearer conversation about platform liability.* Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability

Laslo, M. (2019, August 13). *The fight over section 230 – and the internet as we know it.* Wired. https://www.wired.com/story/fight-over-section-230-internet-as-we-know-it/

Robeznieks, A. (2019, March 19). *Stopping the scourge of social media misinformation on vaccines*. American Medical Association. https://www.ama-assn.org/delivering-care/public-health/stopping-scourge-social-media-misinformation-vaccines

Shane, S. (2017, September 12). Purged Facebook page tied to the Kremlin spread anti-immigration bile. *The New York Times*. https://www.nytimes.com/2017/09/12/us/politics/russia-facebook-election.html

Zittrain, J. L. (2019, October 7). *Three eras of digital governance*. SSRN. https://dx.doi.org/10.2139/ssrn.3458435

Zuckerberg, M. (2018, November 15). *A blueprint for content governance and enforcement.* Facebook. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/