Applying Multi-Stakeholder Internet Governance to Online Content Management

R Street

Free markets. Real solutions.

Applying Multi-Stakeholder Internet Governance to Online Content Management

Chris Riley and David Morar

Table of Contents

The Knight Foundation has provided support to the R Street Institute (R Street) to convene a diverse range of stakeholders to collaborate on problems at the intersection of harm arising from online content moderation, free expression, and online management policies and practices. These are topics that are currently the subject of complex and challenging political discussions in the United States and around the world. With the Knight Foundation's support, R Street seeks both to improve online trust directly through developing a shared understanding and execution of content management practices, and by providing valuable input into ongoing legislative and regulatory conversations regarding the effective role of government.

This report outlines R Street's research that directly explores the complexities of online content management. The report has two distinct and complementary goals: to describe the data collection exercise and to outline the results. The elements of consensus identified in this report aim to ground future conversations in a shared understanding, and the proposals for further attention aim to unlock more granular conversations at the frontier of product and public policy issues and identify opportunities for potential improvement or, in some cases, targeted regulatory intervention.

INTRODUCTION

Online content management is one of the thorniest challenges facing the internet community. Against the backdrop of substantively complex policy debates, political processes often become mired in false binaries in which credentialed experts talk past each other. Often in such tense situations, progress is made only by abstracting problems from their context in order to define and move forward with shared principles and values. While there is benefit in determining shared principles, it remains challenging to return to actionable proposals that risk collapsing delicate values-based alliances. To bolster such efforts, this project sets out to identify a set of concrete and specific intellectual buttresses for further discussion, including proposals that are the subject of active debate, by exploring specific challenges, opportunities, and ambiguities. The hope is that platform managers' and policymakers' future actions will benefit from greater insight into the challenges and opportunities associated with content moderation and recommendation. To ensure clarity and objectivity, the information provided in this report was gathered via a process designed to operate at arm's length from any specific legislative or regulatory contexts, and from the mechanics of trust and safety practices within industry.

A standard joke about politics is that it takes place in smoke-filled rooms, implying that political processes occur among a cloistered few who obfuscate their decision-making in a haze of private dealings. While this image may no longer prevail in the social imagination, it speaks to the undermining of public trust in both the government and in the private sector.[1] Sustainable progress in online trust requires more than unilateral actions by any party; it needs constructive discussions in the open.

As both private sector platforms and public sector government actors evaluate their options for engagement to see whether changes within their power will make things better, it is critical to educate them on the perspectives of all stakeholders involved in the internet's economy and society in order to mitigate any unintended harmful consequences of a particular action or inaction. Among other benefits, such activity will help focus and improve the quality of potential future regulatory or legislative intervention, including by identifying specific issues and questions that could be the subject of constructive future multi-stakeholder efforts. Increasing mutual understanding of online content issues beyond current knowledge by even a single layer of granularity offers the potential for significant benefit.

---

1 See, e.g., "Americans' Views of Government: Low Trust, but Some Positive Performance Ratings," Pew Research Center, Sep. 14, 2020. https://www.pewresearch.org/politics/2020/09/14/americans-views-of-government-low-trust-but-some-positive-performance-ratings; "Techlash? America's Growing Concern With Major Technology Companies," Knight Foundation, March 11, 2020. https://knightfoundation.org/reports/techlash-americas-growing-concern-with-major-technology-companies.

THE CONTEXT

There are many ongoing efforts to study and recommend improvements for content moderation practices, some that began long before or span a broader scale than this report. The value of this report derives from the unique perspective it offers through its mechanisms of inclusive multi-stakeholder participation and through the design of the report itself, which aims toward a more granular educative purpose. Understanding the value of this particular project requires understanding the context in which it takes place, including recent academic scholarship on the range of content moderation remedies and new industry associations.

While academic literature on content moderation spans a far wider range than this report proposes to engage with, we highlight one study representative of the larger body of work due to its representation of the breadth of possible industry actions to mitigate harm online. This study on content moderation articulates 36 separate remedy options across five categories, including content regulation, account regulation, visibility reductions, monetary and "other."[2] The paper includes examples of each of the remedies as used in practice, though of necessity does not provide an individualized analysis of the potential benefits and implementation challenges associated with all of them. Instead, it describes how the remedies can be combined in practice, what factors providers consider when determining which to apply and which relevant cross-cutting factors are involved.

Apart from scholarly work, two trade associations have been created to enrich the landscape of content management practices and discussions. The first, the Trust & Safety Professional Association (TSPA), was established in 2020 and its inaugural Executive Director Charlotte Willner began work in January 2021. The TSPA was created as a hub for practicing trust and safety employees at private companies who are charged with the yeoman's work of protecting online users from harm through applying company policies like initiating content and account takedowns. The trust and safety field is a growing segment of the private sector that is structurally related to the field of privacy. The privacy field has given rise to distinguished institutions of professionalization, notably the International Association of Privacy Professionals (IAPP), which have developed best practices and other mechanisms for elevating the field as a whole; the TSPA seeks to fill a similar need within the field of trust and safety.

Separately, the Digital Trust and Safety Partnership (DTSP), established in 2021, is also structured as a 501(c)(6) membership-led industry association to improve trust and safety practices going forward. Initial signs indicate that the DTSP will embrace a more externally-oriented role, including engagement with public policy processes and government officials, in contrast to the TSPA's more practice-oriented focus.

---

2 Eric Goldman, "Content Moderation Remedies," *Michigan Technology Law Review*, forthcoming March 24, 2021. http://dx.doi.org/10.2139/ssrn.3810580.

While both the TSPA and the DTSP seek to bring in a broad range of stakeholders in their various efforts, neither fits the traditional concept of "multi-stakeholder" as that term is used in the context of internet governance. Multi-stakeholder governance as a paradigm requires the inclusion of perspectives from industry, civil society and government voices in governance discussions. A multi-stakeholder approach to governance begins with the assumption that each group is not just accepted at the table, but encouraged to participate on an even footing and with an equitable place at the table. It is the use of this approach to collaborative ideation and problem-solving for the difficult space of trust and safety that this report seeks to develop.

THE PROCESS

The original stated mission of the multi-stakeholder convening was to begin tackling the global issue of content moderation online at scale in a way that would lead to industry-wide consensus and consensus among stakeholders about fundamental issues. While there was a need for brainstorming and testing possible solutions, the quest for solutions required some shared fundamental understandings and goals, which, given the discourse surrounding social media and online platforms in general, were certainly lacking. The assembly was initially tasked with something far removed from finding solutions to a specific problem. Rather, the goal was to create a set or framework of voluntary industry standards or actions through spirited but collegial debate. The objective at the outset was not to "solve" the issue of online content management, which is an unfeasible objective, but to generate a space for discussion and forward-thinking solutions. While the COVID-19 pandemic upended this convening, the original process was developed using academic literature and empirical examples, and was designed to maximize legitimacy, representation, and diversity in a small-group setting.

The original stakeholder groups were government (both executive and legislative), industry (companies and trade/industry groups), civil society (NGOs, academics, advocates and activists) and journalists. Inviting journalists in as stakeholders allowed for a critical perspective that drew from the personal experiences of those journalists within the process, resulting in rich, fair and knowledgeable reporting. The other groups were seen as standard choices for inclusion in the process. The selection of participants occurred in two waves; first, organizers selected participants with a focus on diversity, then others were given an opportunity to self-select (pending approval by peers in the stakeholder group, and pending veto from the organizers with a written explanation) before the first meeting. After the first meeting each stakeholder group would be instructed to self-organize and create their own structures to internally communicate.

The structure of the convening aimed to circulate around four working groups, with members choosing a primary and secondary working group and the conveners keeping an eye out for proper representation within the groups. Each working group would be focusing on a particular issue, discussed and chosen during the first meeting, with most of the work being done asynchronously through an email listserv. Five day-long meetings would serve as check-in points, allowing face-to-face deliberations and input from participants who were not part of specific working groups. Fundamentally, the participants were set to be in charge of both the substance and type of output that would come out of the working groups, with limited, and light-touch facilitation by the conveners.

Once the pandemic stalled the process, logistical changes crystalized the need for a streamlined, online-only smaller convening with a feasible goal. The process was simplified to a staggered approach, where a small cohort of experts from civil society

(nonprofits, academia, advocacy) and industry (big tech, industry groups) gathered virtually twice to discuss prompts and written work from the conveners, with several one-on-one meetings in between. These changes meant a focus on a smaller group of mostly policy-oriented people, to avoid duplicating trust and safety focused efforts. Doing so shifted responsibility for area of focus and output to the conveners in order to alleviate pressure and gain deep contributions and feedback rather than shared drafting. The conveners also made a concerted effort to capture participants' collective views on the landscape of content management rather than on company-specific issues.

A further change involved reducing the role played by government stakeholders in the process. Government is an integral part of policy, and, taken narrowly, governmental change can be seen as ultimate goal of the policy process. In a multi-stakeholder process, however, government actors are simply one of the participants. The impact of real-world crises and subsequent streamlining of our process meant distilling everything to its core, including the fundamentals of what this process should yield. Multi-stakeholder convenings do not necessarily require all stakeholders to participate, and removing government participants enabled a shift away from the complexities of legislation and toward the fundaments of the issues.

The first meeting took place on Jan. 29, 2021, over Zoom, with two conveners and 13 other participants consisting of six individuals from industry organizations and seven from civil society and academia. The meeting ran for a little over an hour. As with all multi-stakeholder endeavors, getting on the same page with the majority of the participants was a challenge, since scoping the issue required majority buy-in and different points of view yield different perspectives. The conveners allowed the discussion to grow organically, but kept focus broadly on the problem statement. Post-meeting the strategy was adjusted to extract several general propositions and gather individual input, which informed the emergent proposals for further consideration, the final version of which is below.

The initial group was then slightly expanded to include more people who provided feedback on the propositions and general content. While still a closed process at this stage, there was ample space for input and critique on every part of the written text from this larger group via several avenues including email, anonymous feedback and registered (but private) feedback.

A second meeting, on April 6, 2021 gathered the extended group, and focused on seeking specific feedback on the written work assembled by the conveners based on input received thus far. This yielded an initial publicly shared draft. The feedback process then continued asynchronously with the group, and was expanded to include more people through the launch of a public-facing comments-enabled page on the R Street website. While not part of the original version of this multi-stakeholder process, the conveners took their inspiration for this phase from a multi-stakeholder convening in the broader

internet governance space: the NETMundial meeting, wherein an open online repository allowed for comments and feedback from the community on the final text of the statement.[3]

While the conversations leading up to and after the virtual meetings demonstrated an appetite to engage with novel policy making processes around these issues, this translated into fleeting, at best, organic engagement. A potential answer for this conundrum may be that while the funder, institutional home and the people staffing the convening all had good standing in the community, participants found it difficult to make the case that extensive engagement is worthwhile without any initial commitment of shaping future legislation or changing industry or company policies.

To generate more participation the process was extended with additional and more traditional opportunities for input, including a standard Google form and two public virtual webinars with a live question and answer session, which were used both to implement the project's objective of raising awareness, and to promote the opportunity for public input via the form. The two public events, which focused on 1) individual freedoms and protection and 2) industry and ecosystem consequences respectively, were held with experts, drawing from those included in the original group and from new voices, who commented on and discussed the specific propositions from the published draft. Meanwhile, the form was circulated across the technology policy, civil liberties and civil rights communities through a variety of channels and approaches. While, the quantity of engagement remained low, the quality of the input remained high, thus allowing the output of the process as a whole to reflect original and substantial analysis and realize the project's overall goal of contributing meaningfully to broader corporate and policy debates.

3 "NETmundial Comments," NETmundial, 2014, last accessed Aug. 9, 2021. https://document.netmundial.br.

THE OUTPUT

This section reflects the contributions of participants throughout the process as objectively as possible. The below perspectives are not representative of R Street or of the individual authors of this report. Rather, these opinions are presented as an accurate recording from a diverse group of stakeholders drawn from civil society, academia and the tech industry, who participated in this process with no compensation.

A NOTE ON SCOPING

The project is by nature open and inclusive, and took its cues from the participants, both in group conversations and in one-on-one discussions. The initial framing and potential alternatives for the project were proposed and discussed. Any attempt to define a project with specificity inevitably excludes aspects that would be included in other formats, and this project is no exception. Before determining that this project would focus on specific areas of potential further attention, the participants evaluated a broad range of scoping questions and challenges. In particular, they examined:

- Creating specific categories of services for clarity around potential public policy, including how to approach services at different levels of the technical stack.
- Defining the cognizable scope of "harm" and, correspondingly, "harm mitigation" activities.
- How content policies and terms of service are created, and how their enforcement relates to harm mitigation.
- Articulating the broad universe of current practices in content moderation, and in particular the space between "leave up" and "take down" (noting that some other efforts are working in similar directions).
- Collaborating with/ building on existing civil society work to establish a shared glossary of terms.
- Focusing on input from smaller companies and civil society organizations who work with victims of harm as seen by some as underrepresented perspectives in content policy conversations.

This report will explore some of these elements in examining the propositions below, others will be left for future work, and some are likely issues that a multi-stakeholder group will not be able to tackle. The hope is that beyond the final output of this project, the process itself spawns potential future collaborations, a better understanding of the entire ecosystem, new insights into the perspective of other stakeholders and a path forward for action.

Of particular importance to the stakeholders whose input shaped this process is the recognition that this work, like the space of content management more broadly, is not

meant to address the full depth of harm in human connection and communication over the internet. Too often, content moderation is seen as the entire problem and solution for disinformation and hate speech, when it is not. We must all explore potential improvements to day-to-day of online platform practices, while at the same time invest in greater diversity, localism, trust, agency, safety and many other elements. Likewise, content moderation is not a substitute solution to address harms arising in the contexts of privacy or competition.

As always, context is critical, and two directions of evolution of the ecosystem surfaced throughout this process: growing professionalization of trust and safety and moderation processes broadly, and a diversity of thought, design and implementation in these processes. Professionalization can result in a general industry trend of building in better practices from the start, including into business models themselves. At the same time, it is important not to push for content homogenization, but to invest in diversity and experimentation including through consideration of appropriate process rules, transparency and appeals.

POINTS OF CONSENSUS

Identifying points of consensus was not the primary objective of this exercise, but was a consequence of the process itself. Below are broadly shared perspectives that emerged:

- The standards/ expectations for successful content management must not be the perfect and total prevention of online harm, as that is impossible.
- Similarly, content management does not resolve deeper challenges of hatred and harm, and at best works to reduce the use of internet-connected services as vectors
- Automation has a positive role to play in content moderation, but is not a complete solution.
- Automation carries its own risks for internet users' rights, including rights to privacy, free expression and freedom from discrimination.

PROPOSITIONS FOR AREAS OF FURTHER ATTENTION

Each of the following seven propositions represents a specific area that could potentially receive more attention from stakeholders in the content ecosystem, including from industry, civil society, academia and government. Below each proposition are sets of associated positives, challenges and ambiguities provided by participants in this process. These propositions are not presented in any particular order, nor are they sorted according to any intentional method, automated or otherwise.

These propositions are far from comprehensive. In comments submitted to this process, additional ideas were raised that are worthy of broader discussion. For example, one granular idea is to consider collections of posts that collectively result in harm through escalating context, even if each individual item is perhaps not cognizable for moderation. A second broader notion is to dig into user accountability more directly, including through identity-related signaling, such as using privacy-preserving technologies to associate accounts with individual humans (for repeat or duplicate account management) or bots (for alternative processes/ handling). As one commenter stated, "Platforms should know their audience," which is a sentiment particularly true in the context of services used by children and teens. A third suggestion involves providing third-party auditors with access to private data for verification of consistent policy implementation. While there is some overlap between these and the propositions below, these suggestions can also stand alone and could readily be included in future discussions beyond this process.

Proposition 1: Down-ranking and Other Alternatives to Content Removal

The first proposed area of further attention is the use of alternative methods of mitigation for content or accounts in violation of an online service's policies, beyond a

full removal or block. Of particular interest is the use of "down-ranking," which means changing the priority by which content is presented either in response to intentional searching or in recommendation or presentation algorithms. The result is continued accessibility but with reduced visibility.

1. **Positives**

    a. Allows providers to maintain legal speech, but limit its virality, which is a nice compromise.

    b. Can help significantly with non-organic content (e.g., bots and other non-human content contributors).

    c. "Time outs" and temporary limits on sharing of problematic posts or temporary quarantining can help without requiring full removal or blocking of content.

    d. Labeling of content to serve varying purposes, such as to identify partisan political content, bot activity, or intentionally manipulated imagery, can add value.

    e. While scale makes careful consideration of context impossible, other things, like prioritization in terms of account reach, can be done, as well as tiered responses, and other actions that move the conversation away from simply saying context is impossible to deal with at scale.

2. **Challenges**

    a. Can create feelings of being "shadowbanned," or having individuals' reach limited, even where no such intentional activity is occurring, given that there are many factors influencing why a post might not be shown to as many people, including organic signals that the content might not be of broad interest. Transparency and disclosure can help balance that risk.

    b. Can lead to arguments over counterfactuals that are difficult or impossible to disprove, (e.g., "if this content had not been down-ranked, my post would have received X views (and/ or Y dollars)") which can bog down discussion.

    c. In the context of U.S. legal mandates, mandating or incentivizing demotion faces the same First Amendment scrutiny as mandating or incentivizing deletion

3. Ambiguities

a. It is unclear whether down-ranking and similar actions should require notice to the person whose content has been affected. There are arguments on both sides. Traditionally, content rises and falls in search results without any notice as to why; on the other hand, when an explicit intervention has occurred, some due process intuitions point toward giving notice.

b. Seeing and understanding the difference between natural ranking outcomes and intentional actions taken with respect to a specific piece of content (which could be by a company or by a community depending on policy) is not always obvious.

Proposition 2: Granular/ Individualized Notice to Users of Policy Violations

The second proposed area of attention is an increase in granularity and detail in the provision of individualized notices to users whose accounts or content are affected by mitigation methods triggered by policy violations.

1. Positives

a. Individualized, specific notice is arguably a necessary precondition to any due process rights, however minimal.

b. One step lighter is possible: public transparency that lets end users or researchers spot the removal.

c. Content moderation is not just for penalizing bad actors but also educating users on proper use of the system, and individualized notice can greatly assist that educational aspect. This is particularly true with teens and newer users of online services.

d. Clarity of policy in application can reduce confusion and perception of bias

2. Challenges

a. Risks turning everyday content policy disputes into a lengthy process. For example, debate between a platform and a poster over why content was removed is time-consuming and does not often result in mutual understanding.

b. Informing the poster can occasionally agitate the person who reported content due to fear of retaliation.

c. Terms of Service/ Community Guidelines (TOS/CGs) are written broadly and permit case-by-case interpretation. Service providers are always evolving their internal policies and processes in response to new cases. Explaining a decision granularly can create a future expectation of similar result, which should not be guaranteed. This can have the practical effect of redrafting the TOS/ CGs for each decision communicated.

d. Some violations are so voluminous that it would be overly burdensome to notify, or could tip off bad actors to enforcement techniques (for example, commercial spammers).

e. Adds operational burden that is tolerable to incumbents, and less so for their smaller competitors. Every incremental improvement in process requirements is a win for improved content moderation, but likely also a loss for economic viability, resistance to acquisition, etc. of competitors to today's incumbents

3. Ambiguities

a. Everyone has a different idea of what "adequate explanation" of TOS rules or resolution of a particular dispute means in practice. For example, one could ask whether the French civil code's description of "hate speech" is adequate.

b. All users may not be equally entitled to notice, as in cases where notice may help a bad actor, or where bots or people running botnets are the targeted account.

c. An alternative could be more individualized and repeated notice of general community guidelines, outside the context of individual incidents, to ensure expectations are seen and internalized by service users.

Proposition 3: Use of Automation to Detect and Make Classifications of Policy Content

The third proposed area for further attention looks at the use of automation, including context filters of various forms and machine-learning techniques, to evaluate content transactions and detect potential policy violations in real-time. In practice, automation is widely used in real time by many platforms, including for content filtering and

ranking. The extent to which smaller platforms should invest more in automation, or whether more or less automation is desirable compared to the status quo, is the focus of consideration below.

1. Positives

a. Automation can be fast and potentially cheaper, if applicable. It allows application of policies at scale far beyond human moderation.

b. In instances where harms are likely to happen quickly and become high intensity, automated intervention can act as a virality "circuit breaker" and the automated decision can later be modified with more considered thought. For example, after a shooting or during a riot, if accuracy rate is high, it can prevent misinformation from spiraling out of control or escalating crowd behavior.

c. Automated detection paired with human decision-making can make the work more efficient (and in some cases more satisfying). Automation does not need to be all-or-nothing, given the ability to escalate borderline cases to humans.

2. Challenges

a. Automation does a poor job of interpreting context and it is hard to insert context without significant human oversight.

b. Automation costs money, ongoing engineering oversight and support.

c. Errors have disparate impacts, such as with racial bias in hate speech detection.[4]

d. Inserting human oversight will not cure the over-removal problem (because of risk aversion and rubber-stamping) or the disparate impact problem (because of human bias). And it makes the competition problem worse by adding a major additional labor cost.

3. Ambiguities

a. Automation creates a different profile for the articulation of moderation criteria, and some amount of specific tailoring and design is required.

---

4 Maarten Sap, et. al., "The Risk of Racial Bias in Hate Speech Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Association for Computation Linguistics, 2019), pp. 1668-1678. https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf.

b. Automation creates a different profile for errors, including for false-positive and false-negative errors.

Proposition 4: Clarity and Specificity in Content Policies to Improve Predictability at the Cost of Flexibility

Where Proposition 2 proposes increased granularity in individualized ex post notices of policy violation, Proposition 4 proposes increased specificity and detail in the generalized statements of content policy themselves. Although related, the two are different and present different frameworks of analysis and associated incentives and impact.

1. Positives

a. Helps minimize inequitable enforcement by providing predictive clarity.

b. Helps moderation/ policy implementation at scale.

c. Can help automation be more effective.

2. Challenges

a. Does not adapt well to novel circumstances, e.g. does not necessarily provide the flexibility to moderate unanticipated harms.

b. Specific policy text can introduce specific biases.

c. This potentially introduces wider issues related to language, interpretation and analogical reasoning—rules will never be specific or numerous enough to satisfy some. By enumerating certain examples or case studies, there is a risk of enshrining those and suggesting that others are less important. Aspiring for clarity is good but this framing inaccurately suggests that it is possible to write rules specific and comprehensive enough to avoid having to engage in interpretation. It also risks suggesting that a perceived lack of consistency in outcomes is a function of the rule set's design as opposed to a function of variability in human subjectivity.

d. Can encourage narrowing of expectations for good behavior online.

3. Ambiguities

a. Clarity and specificity are good and do not have to take away from flexibility—communication is key because we do not want people to think that everything not specifically banned is necessarily allowed; some catch-all policy is still needed.

b. There is a huge difference between 1) guidance to services on how to develop good policies and 2) what could plausibly/ effectively be in regulatory contexts.

c. Focusing on areas of particular concern to the platform may provide more value, and can vary without changes to underlying content policies.

d. Better reporting tools and improved visibility into what sorts of violative content are identified and acted upon may provide clarity without changing rules.

Proposition 5: Friction in the Process of Communication at Varying Stages

Many of the most cutting-edge experiments in improving the quality of discourse online involve the intentional introduction of friction into communications pathways designed in general to be as frictionless as possible. For example, automated sharing or repurposing of content can be paused, or interstitial pop-ups or other interfaces can be added to the normal user experience flow to prompt for them for more input. Among other variables, such mitigation can be generally applied, temporally implemented during specific offline circumstances or contextually applied where automated mechanisms detect possibly violative content or other triggers of note.

1. Positives

a. Tends to be the most holistic solution and has the advantage of often preventing the bad content before it appears.

b. Encouraging thought pauses before communicating reduces the spread of misinformation and harm.

c. Measuring product interactions (which friction influences) is easier than measuring interpersonal effects.

2. **Challenges**

   a. To the extent that it reduces impulse-clicking and watching, it also reduces ad revenue. Understandably, some other commenters in this process did not view this as a challenge per se, and regard reduced engagement as a positive.

   b. Some users and observers may react badly to friction and view it as a form of covert and therefore dishonest manipulation (as some people view nudges in general).

   c. Measuring harm reduction of friction versus cost to user experience is nontrivial, so optimizing is challenging.

3. **Ambiguities**

   a. It can be tricky to measure the impact, either for discouraging bad actors or bad conduct by well-meaning users.

   b. Discussions of "dark patterns" belong here, and may reveal part of the tension between the challenges and potential positives of reduced engagement.

   c. Scale can make it very difficult to build in things like friction, or user agency perspectives, to potentially create more thoughtful interactions, but that should not mean that the status quo should remain simple, streamlined and fully convenient, which inadvertently may have led to more harm online.

Proposition 6: Experimentation and Transparency in Recommendation Engine Weightings

Related to but distinct from the introduction of friction as part of the user-facing communications flow is the modification of back-end recommendation engines and presentation algorithms, which are used as a means of mitigating online harm, although the details are not always visible to end users. Often such techniques work to combine some of those noted above, including the use of automation (Proposition 3) to engage in down-ranking (Proposition 1); however, as a category, tweaks to the many weighting factors used to determine presentation order for content can go further than these concepts and so it remains interesting as a stand-alone proposition. Furthermore, such weightings can come either from a centralized source or from community sources where decentralized methods shape the presentation of content and/or users.

1. **Positives**

    a. Past examples have shown that it is an effective form of decreasing harm.[5]

    b. It is naturally iterative and responsive to changes in the nature of harm/ impact

    c. Providing more user choice in weighting would increase trust.

2. **Challenges**

    a. Disclosed information re: weighting risks creating genuine confusion among the public.

    b. There is potential risk that adversaries (such as opponents of the tech industry in other industries or in politics) could use it disingenuously.

    c. Without immunity or safe harbor protections, disclosure of weighting could provide fuel for litigation, including adding practical cost to defend against otherwise unsupportable and unjustified suits.

3. **Ambiguities**

    a. Who is doing the weighting? There are significant differences in implications if such considerations are being undertaken by community/users versus the platform.

    b. Competing goals come into focus here, such as European Union lawmakers who want both the promotion of authoritative sources and to ensure diverse perspectives.

    c. Does the underlying recommendation/presentation system use weighting of factors in a way that greater customization makes sense? This may not be something that can be controlled in many contexts.

---

5 YouTube Team, "Continuing our work to improve recommendations on YouTube," YouTube Official Blog, Jan. 25, 2019. https://blog.youtube/news-and-events/continuing-our-work-to-improve.

Proposition 7: Separate Treatment for Paid or Sponsored Content, such as Reviewing for Heightened Standard

Proposition 7 suggests that service providers should hold different standards for content that potentially violates their policies based on whether the content is organic, paid or sponsored by the speaker. This would include payment for placement or prioritization in various forms. Typically, such a difference would apply a heightened standard of responsibility where money is exchanged. In practice, subscription-based services often carry a "know your customer" expectation of responsibility for service providers, which involves knowing or being able to validate the identity of users to a sufficient degree to promote compliance with the law, for example to ensure they are not individuals on U.S. government sanctions entity lists. This proposition would extend that philosophy.

1. Positives

a. With typical paid content, there is more of a direct relationship because money has to change hands. Attaching duties of care in this scenario is very different than for ordinary social media users.

b. Some harms are either mediated through ads (such as scams), or might become unlawful in an advertising context (such as housing or employment discrimination). This method does not fall into the trap of wanting to get rid of a liability shield for categories of content where the First Amendment means there is little possibility of liability to begin with.

c. Paid and sponsored content is at smaller scale than organic content, which allows for more opportunity for pre-moderation and more control, e.g. over placement.

d. The advertising ecosystem already invests in combating ad fraud and ensuring ad integrity, and has long been a focus area for civil rights laws.

2. Challenges

a. In practice, such distinctions tend to focus on specific types of paid content, e.g. political ads, which adds complexity.

b. Far more speech can fall into the paid content bucket than traditional commercial ads, political campaigns, etc. For example, non-governmental organizations (NGOs) using boosting, paying for higher priority and more displays, for their content.

c. Bad actors will find ways to disguise their remuneration flows.

3. Ambiguities

a. Defining paid content is not as straightforward as it seems. For instance, in a freemium or subscription model, does all content become paid? What about individual users organically giving digital awards to each other's posts? What does "paid" content mean for hosting services? Turning principle into practice requires clarification of such distinctions for precise application.

b. A threshold could be used for further reduction of scope, for example to put small-dollar transactions into the same category of protection as organic content.

NEXT STEPS

This report will be made public and shared intentionally with key stakeholders, including policymakers within the United States and the European Union. Ideally, the report's granular discussions of frontier policy questions in the online content space will improve shared understanding on critical issues.

This process, conducted under the auspices of R Street Institute, produced original and substantial contributions to the ongoing dialogue. However, structural limitations in its scope resulted in high quality but low quantity engagement. In contrast, multi-stakeholder processes that are convened by national governments build on their inherent legitimacy, substantial resources, and clear potential for shaping future regulatory outcomes to drive sustained and scaled engagement. For example, the National Institute of Standards and Technology (NIST) conducted an extensive multi-stakeholder process on cybersecurity resulting in a framework of standards, guidelines and best practices.[6] Likewise, the National Telecommunications and Information Administration (NTIA), also part of the Department of Commerce, conducted a number of multi-stakeholder processes on internet and technology policy topics, such as facial recognition.[7] Little wonder that Secretary Raimondo of the U.S. Department of Commerce indicated in her confirmation hearing that she intends ask NTIA to "convene stakeholders" to work on this issue.[8]

An ideal next step for the strategic and substantive discussions teed up by this project would be the commencement of a more extensive multi-stakeholder discussion on online content management—perhaps via NTIA—with the goal of building on the key propositions identified through this process as well as others that arise through broader engagement.

---

6 "Cybersecurity Framework," National Institute of Standards and Technology, last accessed Aug. 31, 2021. https://www.nist.gov/cyberframework.

7 "Privacy Multistakeholder Process: Facial Recognition Technology," National Telecommunications and Information Administration, June 17, 2016. https://www.ntia.doc.gov/other-publication/2016/privacy-multistakeholder-process-facial-recognition-technology.

8 Emily Birnbaum, "Commerce Department nominee advocates for Section 230 reform," *Protocol*, Jan. 26, 2021. https://www.protocol.com/bulletins/gina-raimondo-section-230-reform.