
Beyond Content-Based Enforcement

Jeffrey W. Howard

Behavior-based interventions involve restricting users based on what they do, rather than what they post. This practice is already established at major platforms. However, it has received little explicit philosophical scrutiny. This essay explores behavior-based interventions from the perspective of ethics and political philosophy. It first distinguishes interventions targeting behavior that is itself a rule violation (such as impersonation) from the harder class of cases in which no single action breaches a rule, but a combination of signals indicates misuse. It then argues that, because such interventions are best understood as defensive, the philosophical literature on defensive ethics yields the relevant criteria they must satisfy. Interventions should respect user autonomy by being triggered by conduct for which the user is responsible, ruling out demographic profiling and purely association-based signals. They should be suitable, necessary, and proportionate, with particular attention to the risk of false positives. The essay also defends the importance of transparency, graduated interventions, and meaningful appeals mechanisms. It flags the distinctive danger of governments seeking platforms' behavioral predictions. And it closes by urging platforms to institutionalize engagement with scholars, civil society, and affected rights-holders.

Consider an example of an adult user on a social media platform attempting to connect with 13-year-olds with whom he has no prior connection. He hasn't violated any rules on violating content in these conversations—for example, sending or soliciting nude photos—at least, he hasn't violated them yet. Still, the behavior is suspicious. What should the platform do? Should it, for example, limit the ability of this person to contact minors in the future, to lower the risk of harm? Should it only do so if it has some further signal to bolster the suspicion—for example, if the user has also engaged in other suspicious behavior, such as searching for terms related to child sexual abuse material (CSAM) or leaving sexualized comments on images of minors?

For much of the past decade, conversations around digital platforms with centralized moderation have centered on the rules determining what users can and cannot post online. This often leads to binary debates over whether to allow certain types of speech.

However, stakeholders across industry (Pugalia and Lalani 2023), civil society (Collier 2021), and academia (Iyer 2022) often highlight how content-based enforcement, though effective in tackling one-off violations, may be insufficient in addressing the more persistent and systematic misuses of platforms. Given the scale at which these platforms moderate, focusing on content in isolation may seem like a limited solution, placing moderators one step behind users determined to evade the rules. Moreover, many misuses of platforms don't involve posting prohibited content, such as foreign influence campaigns posting divisive but otherwise permitted speech. Some misuses don't even involve posting anything (e.g., creating impersonating accounts).

To address these gaps, platforms can try to reach upstream from content and instead assess *behaviors* that correlate with platform misuse. This includes such behaviors as who people are messaging, how many times they message, what forums or channels they follow, what types of information they're seeking out, etc. The leveraging of these signals for purposes of moderation is what is often called *behavior-based intervention*. The term and the practice are not my invention: as I discuss below, such interventions are already well established at major platforms. My aim is rather to subject them to more explicit normative scrutiny.

Such intervention is appealing for a couple of reasons. One is that it can be effective in removing accounts of people who can be confidently assessed to be violating platform policies; at least for the most severe violations, this should be relatively uncontroversial. Focusing on these accounts can also have an outsized impact on the prevalence of violating content, given that the vast majority of people, by contrast, obey platform rules. Finally, and perhaps counterintuitively, this approach to enforcement may safeguard freedom of expression by embracing a more targeted approach to centralized moderation—one that reduces harm while removing less non-violating content.

Today, there is growing interest in applying this approach across violation types. Yet such measures raise important questions. How can platforms adequately balance important principles such as accountability, privacy, fairness, reliability, safety, transparency, and security when deploying this type of online moderation? And how can they strike that balance at the scale at which enforcement actually operates, where decisions are made in bulk by a mixture of automated systems and human reviewers, and where trade-offs are inevitable?

My aim here is to explain the benefits, challenges, and trade-offs associated with behavior-based intervention, and to consider how platforms can develop appropriate normative frameworks by drawing on a rich body of existing scholarship related to political philosophy, criminal justice, and human rights.

1 What Are Behavior-Based Interventions? A Closer Look

The type of moderation people tend to think of first is *content-based*; users are sanctioned for posting content that violates the platform's rules. Behavior-based interventions are different—it's what users do, not what they say. Of course, posting prohibited user-generated content is itself a behavior. So *behavior-based* intervention is a term of art, referring to interventions that do not fit the paradigmatic content-based case.

What kind of behavior could trigger intervention? Conceptually, we can distinguish between two different kinds. In one set, *intervention directly targets violating behavior*. For example, it is a violation of Meta's Community Standard on Authentic Identity Representation to create an account impersonating another person, or to adopt a fake identity (Meta, n.d.-b). If Meta confirms that an account is an impersonation, it will be removed. In so doing, Meta is straightforwardly enforcing its rules, as when it removes content violating its other Community Standards.

But there is another class of behaviors that could also trigger intervention. These behaviors, by themselves, are not rule violations. Rather, these are a combination of behaviors that can be *assessed* as probable rule violations. Consider an example of an adult attempting to connect with a minor with whom they have no prior connection. Perhaps that adult is a new coach or teacher, but maybe they've also engaged in concerning conduct in the past, searching for terms related to child sexual abuse material (CSAM) or leaving sexualized comments on images of minors. Because this latter behavior may be indicative of future misuse, it could be a basis for intervention.

This kind of *behavior-based intervention* is an important but under-discussed element of platform governance. One example is Snapchat's use of sextortion behavior signals to identify and remove bad actors before they target and victimize others (Snap Inc. 2024). Another is Google's use of actor and behavioral signals to combat scams (Google 2024). A third is Meta's efforts to identify potentially suspicious adults on Facebook and Instagram, and then prevent them from discovering and connecting with teen accounts (Davis 2021). In conjuring up these policies, platforms developed a typology of behavioral signals they have found to correlate with certain misuses of their platforms. They then identified a threshold at which they have sufficient confidence to intervene, which may take the form of limiting a user's access to others or even disabling a user's account.

There are numerous ways in which this approach could be effective. One consistent finding from research on online behaviors is that negative or disruptive behaviors tend to be concentrated among a small number of individuals. For example, a 2018 study of conflict between Reddit communities estimated that 1 percent of subreddit communities initiate 74 percent of all intercommunity conflicts on the platform (Kumar et al. 2018). That finding concerns communities rather than individual users, but the same concentration has been documented at the individual level across a range of harms: studies of misinformation sharing (Grinberg et al. 2019; Guess, Nagler, and Tucker

2019; Allen et al. 2020), engagement with radical content (Hosseinmardi et al. 2021), and online hate (Zannettou et al. 2020) consistently find that a small fraction of users accounts for the vast majority of the behavior—a pattern this journal’s special issue on the topic described as “uncommon yet consequential” online harms (Robertson 2022). This research suggests that a system that consistently and accurately identifies this small percentage of users will have a tremendous impact on maintaining safety online while reducing entanglement with the debates about online speech. These debates often struggle to find consensus on how platforms should define—let alone sanction—speech that is misleading, hateful, or incendiary. While platforms may never escape these decisions and their attendant criticisms, wherever they draw the line, they may be able to simultaneously respect voice and address safety issues by tactically targeting those truly attempting to inflict harm.

Behavior-based interventions are not merely theoretical: platforms already deploy them in relation to content-based harms (e.g., targeting behaviors highly correlated with future posting of violating content). And they are especially well established in combating issues that have more content-agnostic elements—most notably a family of coordinated, often automated, behaviors that most platforms’ terms of service already expressly prohibit.¹ Here the concern is not with any individual piece of content, but rather with whether the user’s behavior is persistent, repetitive, or pursued in coordination with others to pursue an adversarial or prohibited goal. In these cases, targeting patterns indicative of misuse may be the only effective lever platforms have to stymie them.

However, these interventions also raise a wide range of difficult questions. When exactly is it fair to limit a user’s ability to use a digital service not based on their visible posts, but rather on assessments of their conduct? How should such assessments adjust to accommodate local context? Do such interventions reflect a paranoia that platforms are filled with bad-faith actors and what would that mean for the usability of the platform? How do you simultaneously protect the ability of good-faith actors to access information, associate with each other, and build community while also protecting them and others off-platform from potential harms? Are these interventions even effective in reducing conduct risks, and if so, how would we measure that?

1. These include coordinated inauthentic behavior (networks of inauthentic users seeking to manipulate public debate for a strategic goal); astroturfing (artificially creating the appearance of widespread support for something) (Chan 2024); review-bombing (insincerely leaving negative reviews of certain products to manipulate rankings) (Zaroli 2023); vote manipulation (insincerely upvoting or downvoting authentic content to manipulate rankings) (Carman et al. 2018); and brigading (coordinated efforts to report certain posts as violating) (Andrews 2021). Because such coordinated—and frequently automated—behaviors constitute a specific subclass of behavior-based issues, and are typically already explicit violations of platforms’ terms of service, I set them aside in what follows.

2 The Ethics of Behavior-Based Interventions

I cannot hope to answer all the relevant questions about behavior-based interventions in one article. What I can do—in the hope of demonstrating how platforms can incorporate existing schools of thought into their policy development—is take some of the most pressing questions and couch them in terms of familiar ethical dilemmas. (One question I deliberately do not pursue is how to draw a clean conceptual line between “content” and “conduct”—a task that could occupy an article of its own. I work instead with paradigm cases of the kind described above.)

Many of the aforementioned questions, like many ethical questions, cannot be answered through empirical testing; they are normative questions that require each platform to consider how to strike a balance between different ethical values when they seem to conflict. These are questions that also depend on the digital service in question, each of which possesses unique affordances and vulnerabilities. Below, I try to articulate the best possible case for behavior-based interventions, explaining what it would take for such interventions to be justified. I then tackle how to operationalize some of the proposed ethical underpinnings.

As some scholars have argued, platforms should think of their policies against harmful content and conduct as aimed at defending their users from harms that can occur on their platforms—as well as people off-platform who can be harmed by conduct originating there, as when a scam initiated on a platform defrauds its victim elsewhere (Howard 2024). They are a matter of defense, not retributive punishment. So it makes sense to turn to philosophical insights about the ethics of preventive and defensive intervention, including the ethics of policing. This vast philosophical literature yields a range of important criteria that any intervention must meet to be justified.

2.1 Respecting Users’ Autonomy

First, and perhaps most importantly, an intervention should err toward *respecting users’ autonomy*. That means that it should be triggered by something the user has done, for which they are responsible. Consider a case of defensive harm, in which a gunman is attacking innocents in a park. Is it compatible with respect for the gunman’s autonomy for police to forcibly intervene to stop the attack? Obviously, it is. The reason is that, in such a situation, the gunman’s *own free choices* have made it necessary to intervene to prevent harm to others. He cannot complain because he has brought the situation on himself. He is, as legal scholars and philosophers would say, *liable* to suffer this harm, meaning he is not wronged by it. The intervention respects his autonomy as he could have decided not to bring the situation about in the first place to spare himself harm.

Now, consider a harder case. Suppose the police receive a phone call from a librarian at a university that a student has been searching repeatedly for instructions on how to make an explosive device. This is no guarantee that the student is aiming to break the law; perhaps

their interest is purely intellectual. But if this information is assessed against other particulars (e.g., a report describing the student's erratic behavior and an anonymous threat received by administrators), the combination of signals may make it reasonable for police to be suspicious and even pursue certain interventions—at minimum, questioning the student. Yet it is the student's own choices that make such intervention appropriate. Those choices have made it reasonable for others to grow concerned. Accordingly, they cannot complain if subjected to an intervention (assuming further conditions, discussed below, are also met). Indeed, it can be argued that people have duties to bear certain costs to reassure others that they don't pose a threat (which is why you have a duty to open the trunk of your car when driving into an airport in countries with terrorism risks). The same analysis can be used to justify other forms of policing, such as imposing restraining orders on those deemed dangerous. Each case is different, to be sure, but they are governed by the same underlying normative standards.

Some online behavior is roughly analogous to the library scenario.² Consider again the adult user who persistently reaches out to minors with whom they have no prior connection. This suspicious behavior may make it reasonable for the platform to take action on the basis that they have illicit intentions. And their own decisions may make it reasonable for the platform to suspect this. Or consider the user who, shortly after joining, begins sending near-identical messages to large numbers of strangers with whom they have no connection. To be sure, some versions of this conduct (such as operating networks of fake accounts) are already per se violations of most platforms' rules, requiring no probabilistic inference at all. The trickier cases are those in which no single action breaches a rule, yet the overall pattern is highly correlated with inauthentic behaviors like influence campaigns or scams (Meta, n.d.-d). In such cases, the user's own conduct has made it reasonable for the platform to consider intervention.

Contrast this with a case in which a platform decides, because of a user's demographic profile, that they are more likely than other users to engage in certain platform wrongs. *That* type of intervention is discriminatory and blatantly fails to respect user autonomy. Such profiling makes it functionally wrong to belong to a particular reference class to which a statistical probability is indexed. It fails to treat people as individuals to be held accountable for their wrongful or otherwise suspicious choices, and treats them merely as members of a group. Another non-autonomous signal that veers from the notion of respecting user autonomy would be an intervention based exclusively on how others have interacted with someone (e.g., a user gets sent a lot of violating content or is repeatedly reported) without any indication of their own action. The same worry extends to a broader family of association-based signals. These include intervening against someone because of who follows or messages them, because they belong to a group in which others misbehave, because they were added—through no action of their

2. Though not, importantly, to the park scenario. The gunman is not merely behaving in ways that correlate with future harm; he is already inflicting the very harm to be prevented. His online analogue is the user caught in outright violation (such as actively sending abusive material to a minor), to whom straightforward enforcement applies. The harder cases parallel the student in the library.

own—to a channel later found to be violating, or because adversaries have weaponized reporting tools against them. What unites these scenarios is that the triggering “behavior” consists entirely of other people’s choices. Platforms should treat such signals, at most, as prompts to scrutinize what the user has actually done. Alone, they should never be sufficient grounds for intervention.

To be clear, the difficulty here is not the use of actuarial methods or AI as such: at platform scale, behavior-based intervention inevitably relies on statistical models to aggregate and weigh signals. The difficulty concerns what those models take as inputs. Predictions grounded in a user’s own responsible conduct—what they have chosen to search for, whom they have chosen to contact, how they have chosen to behave—can respect autonomy in the way described above. Predictions grounded in features outside the user’s control—demographic attributes, or how others have treated them—cannot, and using them as the sole basis for restricting someone’s opportunities is very difficult to justify (Jorgensen 2022). At the very least, then, among the signals that trigger an intervention there should be specific behavior for which the user is responsible. Much accordingly rides on which behaviors platforms select as signals, and where the line of responsibility is drawn.

2.2 Assessing Suitability, Necessity, and Proportionality

Even if an intervention errs toward respecting autonomy, that doesn’t mean it is justified. Further conditions should be met. Drawing from jurisprudence, an intervention should be *suitable*, *necessary*, and *proportionate*. For every candidate intervention, we must work through whether it satisfies each criterion. Roughly, an intervention is *suitable* if it stands a reasonable chance of achieving its goal, ruling out interventions that would be manifestly ineffective. An intervention is *necessary* if there is no way of achieving the same goal with an action that is less costly to the target. And an intervention is *proportionate* if its costs (both to the target and to those collaterally burdened by enforcement mistakes) aren’t excessively large compared to the benefits.

There’s no prescriptive playbook for how to apply these tests; what counts as “suitable enough,” “necessary enough,” or “proportionate enough” involves careful consideration for each digital service and the salient risks involved. For example, if an adult user on a large platform is reaching out to minors with whom they have no prior relationship, it may be disproportionate to simply ban them from the platform, and perhaps not strictly necessary to achieve the relevant goal (preventing harmful contact with minors). In other cases, as with suspected fake accounts engaged in behavior that purposefully employs deceptive means, such as the coordinated use of multiple accounts to either scam or defraud users, banning those accounts would likely be both necessary and proportionate. Sometimes, depending on the severity of the risk, the potential for false positives may be too high for intervention to be proportionate, since too many good-faith users would be affected. One can imagine users carrying out research about a topic involving restricted goods or services or small business owners trying to build a clientele

by rapidly connecting with strangers online.

Platforms are, in practice, well aware of this risk. Platforms' own research counsels caution about what such signals actually show. In an analysis accompanying the child-safety efforts noted above, Meta reported that, of 150 accounts it had reported to the National Center for Missing and Exploited Children for uploading child exploitative content, an estimated "more than 75% of these people did not exhibit malicious intent (i.e. did not intend to harm a child)"; instead, they "appeared to share for other reasons, such as outrage or in poor humor"—findings Meta itself stressed "should not be considered a precise measure of the child safety ecosystem" (Davis 2021). The lesson is twofold. First, behavioral signals can surface serious wrongdoing that content review alone would miss. But, second, individual signals are often deeply ambiguous—which is why the combination of signals (i.e., not over-indexing on a single signal), and the confidence threshold at which platforms act, matter so much.

All of these are open and contested questions to be asked on a case-by-case basis. My point is simply to highlight that applying these criteria is essential in striking the right balance between competing values every platform must consider—such as freedom of expression, privacy, the dignity of victims, and the well-being of online communities.

Fortunately, existing frameworks can help platforms properly weigh these trade-offs. Indeed, the concept of proportionality is a well-established principle in international human rights law, which has already come to play a significant role in guiding decision-making and processes by platforms (Levy Daniel and Ross 2026). The lines that do end up being drawn will surely look different depending on the nature of the violation, the severity of potential harm, the accuracy of enforcement tools available, among other factors. Compounding the complexity is the challenge, flagged at the outset, of executing behavior-based enforcement at scale, which inevitably means leaning in favor of some values over others. Platforms have to recognize and tolerate that elasticity, even while striving for continuously improved precision.

2.3 Determining a Reasonable Level of Transparency and Remedy

Once justification is established, other important questions remain over determining how to inform users, retain trust, and ensure enforcement remains effective.

Platforms want to provide a reasonable level of transparency to users about their rules and interventions. However, what counts as reasonable is sensitive to many considerations. One is the cost of intervention. It is not a massive cost to be deprived of the opportunity to reach out to minors with whom one has no prior connections—so must one be notified when they've incurred an intervention? Another is the risk that adversarial actors will use detailed descriptions of violating behaviors to circumvent enforcement. The Russian government would like nothing more than a detailed playbook of how platforms detect their influence operations, precisely so they can evade it. Thus,

there is a trade-off between how much transparency users are afforded, on the one hand, and the overall efficacy of the intervention, on the other.

That trade-off is more subtle than it first appears, as transparency can itself contribute to efficacy. In a survey study of rule violators on Twitter, paired with an analysis of their subsequent platform behavior, users who felt more fairly treated during an enforcement action were less likely to reoffend (Katsaros et al. 2022). Procedural fairness, in other words, is not simply a cost paid at efficacy's expense; sometimes it can be part of what makes enforcement effective. The case for withholding detail is strongest for sophisticated adversarial actors, and weakest for the ordinary users who—as Meta's intent research above suggests—make up many of those swept up by behavioral signals.

At a minimum, platforms should be transparent about the fact that they *have* behavioral interventions in place. But should they spell out publicly what signals they deem to be indicative, and if so, to what extent? For example, Meta's "Account Integrity" policy provides that they will disable accounts that "[d]emonstrate activity or behavior indicative of a clear violating purpose," and even lists a few examples of the signals it uses (Meta, n.d.-a). That question—what counts as reasonable transparency—is bound to be debated in this domain as it is in every other. It will depend on how to strike the balance between users' interests in understanding the rules to which they are subject and their interest in not being subjected to potential harms from another user's conduct.

Here again, existing frameworks can be instructive. In particular, the United Nations Guiding Principles on Business and Human Rights (UNGPR) provides a list of criteria for grievance mechanisms (United Nations 2011). In behavior-based enforcement, the notion of *transparency* suggests that people should be notified if their behavior is likely to trigger an intervention into their digital lives, especially if the costs are great, so they can avoid bearing those costs. Access to remedy means that users should be provided with grievance pathways to report concerns as well as seek remedy and remediation.

The UNGPR's "effectiveness criteria" for grievance mechanisms offer more specific guidance still. A mechanism should be predictable. In behavior-based enforcement, this favors graduated interventions (warnings, feature limits, temporary restrictions) before account disablement wherever the severity of the risk permits, so that users learn where the lines are drawn through the enforcement process itself. Further, a user appealing a behavioral intervention should be told the category of behavior at issue, even where the precise signals remain confidential. And it should be a source of continuous learning: appeal outcomes and reversal rates for behavior-based actions are a measure of false positives, and should be fed back into the calibration of signals and thresholds. Ideally, they should also be disclosed in transparency reports, as civil society frameworks urge for content-based enforcement (Santa Clara Principles 2021). Indeed, enforcement that users experience as fair tends to work better (Katsaros et al. 2022).

Finally, behavior-based enforcement gives a sharper edge to a familiar concern about

government demands for platform data. That platforms hold extensive data on their users' behavior is a longstanding issue, and it arises independently of the interventions discussed here. What is newer is that behavior-based systems generate derived signals—typologies of suspicion, risk scores, predictions of future wrongdoing—that governments may seek for their own purposes. Sharing an inference that a user is likely engaged in some category of wrongdoing raises different, and arguably graver, concerns than sharing a specific piece of content, especially in jurisdictions where the predicted “wrongdoing” is protected expression or association. Platforms deploying behavior-based interventions will therefore need policies that specifically address requests for behavioral signals and predictions, informed by existing guidance (such as the Global Network Initiative's principles) on responding to government demands in a manner consistent with users' rights to freedom of expression and privacy (Global Network Initiative, *n.d.*).

3 Operationalizing Interventions and Governing Deployment

We have only scratched the surface of the complexities surrounding this topic and as is often the case with frameworks that offer compelling justifications for intervention, the devil is in the details when it comes to implementation.

This is a place where platforms can strongly benefit from a practice of direct and sustained engagement with external stakeholders to help adapt frameworks. Through roundtables, workshops, and one-on-one meetings, external engagement can play an instructive role in helping translate high-level principles into operationalized platform governance. Meta has pursued precisely such engagement in developing and refining its Community Standards (Meta, *n.d.-c*). Yet such engagement is only as durable as a platform's commitment to it. Meta's abrupt overhaul of its content policies in January 2025, announced with seemingly little of the consultation that characterized earlier revisions, illustrates that voluntarily adopted processes can be discarded at a platform's discretion. This is a clear limitation of any approach that relies on voluntary engagement, and a reason external stakeholders reasonably press platforms to institutionalize such commitments rather than treat them as revocable gestures. By engaging directly with academia and civil society organizations, including scholars of international human rights law and political philosophy, internal teams can learn from fields of study that, while different in some respects, tackle complexities similar to those posed by behavior-based enforcement.

Ultimately, platforms must decide the threshold at which they are sufficiently confident to intervene, which may take the form of limiting a user's access to others, disabling certain product features, or even disabling a user's account. But along the way, the forms of engagement described above—consultation with scholars, civil society organizations, and affected rights-holders—can help teams develop sound guardrails to mitigate risks, such as bias, which can lead to over-enforcement on benign users or amplify existing shortcomings in platform moderation efforts.

4 Conclusion

Purely content-based enforcement will remain a core feature of platform governance. But behavior-based interventions represent a powerful tool that allow platforms to be more timely and efficient in managing digital spaces. They also raise valid concerns pertaining to user autonomy and more. Beyond voicing their opinions by sharing content, internet users show up to find community, meet new people, and engage with ideas and opinions by searching for, “liking” or “loving” content. Platforms that wish to retain the trust of their users—and, increasingly, to satisfy their regulators—must sustain this openness. Meanwhile, bad actors seek to exploit these same features in ways that can contribute to a variety of online and offline harms.

By borrowing from fields such as political philosophy, platforms can bring added rigor to enforcement practices that already extend well beyond reacting to individual posts. I have tried to show concretely what that borrowing involves: grounding interventions in behavior for which users are responsible; testing them against the criteria of suitability, necessity, and proportionality; treating transparency and remedy not merely as costs but as components of effective enforcement, in line with the UNGP’s effectiveness criteria and the empirical literature on procedural justice; and institutionalizing external engagement so that it survives shifts in platform leadership and priorities. A rich, multidisciplinary conversation on these questions is already under way across academia and civil society. My aim has been to connect the operational choices platforms face in behavior-based enforcement more tightly to that conversation.

References

- Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the Fake News Problem at the Scale of the Information Ecosystem." *Science Advances* 6 (14). <https://doi.org/10.1126/sciadv.aay3539>.
- Andrews, Phoenix C. S. 2021. "Social Media Futures: What Is Brigading?" Tony Blair Institute for Global Change, March 10, 2021. <https://institute.global/insights/tech-and-digitalisation/social-media-futures-what-brigading>.
- Carman, Mark, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. "Manipulating Visibility of Political and Apolitical Threads on Reddit via Score Boosting." In *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications / 12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE)*. <https://ieeexplore.ieee.org/abstract/document/8455905>.
- Chan, Jovy. 2024. "Online Astroturfing: A Problem beyond Disinformation." *Philosophy & Social Criticism* 50 (3): 507–28. <https://doi.org/10.1177/01914537221108467>.
- Collier, Anne. 2021. "Countering 'Hate Raids' Needs to Be Coordinated Too." NetFamilyNews.org, September 2, 2021. <https://www.netfamilynews.org/countering-hate-raids-needs-to-be-coordinated-too>.
- Davis, Antigone. 2021. "Preventing Child Exploitation on Our Apps." Meta Newsroom, February 23, 2021. <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>.
- Global Network Initiative. n.d. "The GNI Principles." Global Network Initiative. Accessed July 27, 2026. <https://globalnetworkinitiative.org/>.
- Google. 2024. *Tackling Scams and Fraud Together*. Google. https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/Tackling_scams_and_fraud_together.pdf.
- Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter during the 2016 U.S. Presidential Election." *Science* 363 (6425): 374–78. <https://doi.org/10.1126/science.aau2706>.
- Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5 (1). <https://doi.org/10.1126/sciadv.aau4586>.
- Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. "Examining the Consumption of Radical Content on YouTube." *Proceedings of the National Academy of Sciences* 118 (32). <https://doi.org/10.1073/pnas.2101967118>.
- Howard, Jeffrey W. 2024. "The Ethics of Social Media: Why Content Moderation Is a Moral Duty." *Journal of Practical Ethics* 11 (2). <https://doi.org/10.3998/jpe.6195>.

- Iyer, Ravi. 2022. "Content Moderation Is a Dead End." *Designing Tomorrow* (Psychology of Technology Institute), Substack, October 7, 2022. <https://psychoftech.substack.com/p/content-moderation-is-a-dead-end>.
- Jorgensen, Renée. 2022. "Algorithms and the Individual in Criminal Law." *Canadian Journal of Philosophy* 52 (1): 61–77. <https://doi.org/10.1017/can.2021.28>.
- Katsaros, Matthew, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. "Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter." *Journal of Online Trust and Safety* 1 (3). <https://doi.org/10.54501/jots.v1i3.38>.
- Kumar, Srijan, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. "Community Interaction and Conflict on the Web." In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, 933–43. <https://doi.org/10.1145/3178876.3186141>.
- Levy Daniel, Maia, and Hilary Ross. 2026. "Trust and Safety and Human Rights: Bridging the Fields for Better Online Governance." In *Trust, Safety, and the Internet We Share: Multistakeholder Insights*, edited by Maia Levy Daniel, Amanda Menking, Marlyn Savio, and Jean Claffey, 104–16. New York: Routledge. <https://doi.org/10.1201/9781003621072-14>.
- Meta. n.d.-a. "Account Integrity." Meta Transparency Center. Facebook Community Standards. Accessed July 27, 2026. <https://transparency.meta.com/policies/community-standards/account-integrity>.
- . n.d.-b. "Authentic Identity Representation." Meta Transparency Center. Facebook Community Standards. Accessed July 27, 2026. <https://transparency.meta.com/policies/community-standards/authentic-identity-representation>.
- . n.d.-c. "Facebook Community Standards." Meta Transparency Center. Accessed July 27, 2026. <https://transparency.meta.com/policies/community-standards/>.
- . n.d.-d. "Inauthentic Behavior." Meta Transparency Center. Facebook Community Standards. Accessed July 27, 2026. <https://transparency.meta.com/policies/community-standards/inauthentic-behavior/>.
- Pugalia, Akash, and Farah Lalani. 2023. "Digital Safety: How a Multi-Faceted Approach Can Help Tackle Real-World Harm." World Economic Forum, September 29, 2023. <https://www.weforum.org/stories/2023/09/digital-safety-multi-faceted-approach-tackle-real-world-harm/>.
- Robertson, Ronald E. 2022. "Uncommon Yet Consequential Online Harms." *Journal of Online Trust and Safety* 1 (3). <https://doi.org/10.54501/jots.v1i3.87>.
- Santa Clara Principles. 2021. "The Santa Clara Principles on Transparency and Accountability in Content Moderation." 2nd edition. <https://santaclaraprinciples.org/>.
- Snap Inc. 2024. "Our Work to Help Keep Snapchatters Safe." Snap Privacy and Safety Hub, October 4, 2024. <https://values.snap.com/news/keeping-snapchatters-safe>.

- United Nations. 2011. *Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework*. HR/PUB/11/04. New York: Office of the United Nations High Commissioner for Human Rights. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinessshr_en.pdf.
- Zannettou, Savvas, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. “Measuring and Characterizing Hate Speech on News Websites.” In *Proceedings of the 12th ACM Conference on Web Science (WebSci ’20)*, 125–34. <https://doi.org/10.1145/3394231.3397902>.
- Zaroli, Jim. 2023. “Goodreads Has a ‘Review Bombing’ Problem — and Wants Its Users to Help Solve It.” NPR, December 17, 2023. <https://www.npr.org/2023/12/17/1219599404/goodreads-review-bombing-cait-corrain>.

Authors

Jeffrey W. Howard is Professor of Political Philosophy and Public Policy at University College London. Email: jeffrey.howard@ucl.ac.uk.

Acknowledgments

I am indebted to Conor Sanchez for countless conversations on this topic, which have taught me much and substantially shaped my views. I am grateful to the editors of this journal for exceptionally constructive feedback. Thanks also to audiences at the Third International Platform Governance Research Network Conference; the AI and the Future of Platforms Conference at the University of Hong Kong; and the Content Moderation Lab at the Technical University of Munich.

Funding Statement

I am thankful to UK Research and Innovation for research funding (UKRI grant MR/V025600/1).

Disclosure Statement

The author is an adviser to the UK Office of Communications.

Keywords

Behavior-based intervention; content moderation; platform enforcement; youth safety; ethical frameworks