Fall 2022

Countering Personalized Speech

Leon G. Ho

Follow this and additional works at: https://scholarlycommons.law.northwestern.edu/njtip

Part of the Communications Law Commons, Entertainment, Arts, and Sports Law Commons, First Amendment Commons, and the Intellectual Property Law Commons

Recommended Citation
https://scholarlycommons.law.northwestern.edu/njtip/vol20/iss1/2

This Article is brought to you for free and open access by Northwestern Pritzker School of Law Scholarly Commons. It has been accepted for inclusion in Northwestern Journal of Technology and Intellectual Property by an authorized editor of Northwestern Pritzker School of Law Scholarly Commons.
COUNTERING PERSONALIZED SPEECH

Leon G. Ho

November 2022

© 2022 by Leon G. Ho
COUNTERING PERSONALIZED SPEECH

Leon G. Ho*

ABSTRACT— Social media platforms use personalization algorithms to make content curation decisions for each end user. These personalized recommendation decisions are essentially speech conveying a platform’s predictions on content relevance for each end user. Yet, they are causing some of the worst problems on the internet. First, they facilitate the precipitous spread of mis- and disinformation by exploiting the very same biases and insecurities that drive end user engagement with such content. Second, they exacerbate social media addiction and related mental health harms by leveraging users’ affective needs to drive engagement to greater and greater heights. Lastly, they erode end user privacy and autonomy as both sources and incentives for data collection.

As with any harmful speech, the solution is often counterspeech. Free speech jurisprudence considers counterspeech the most speech-protective weapon to combat false or harmful speech. Thus, to combat problematic recommendation decisions, social media platforms, policymakers, and other stakeholders should embolden end users to use counterspeech to reduce the harmful effects of platform personalization.

One way to implement this solution is through end user personalization inputs. These inputs reflect end user expression about a platform’s recommendation decisions. However, industry-standard personalization inputs are failing to provide effective countermeasures against problematic recommendation decisions. On most, if not all, major social media platforms, the existing inputs confer limited ex post control over the platform’s recommendation decisions. In order for end user personalization to achieve the promise of counterspeech, I make several proposals along key regulatory modalities, including revising the architecture of personalization inputs to confer robust ex ante capabilities that filter by content type and characteristics.

* J.D., University of North Carolina School of Law. I thank the faculty at the University of North Carolina School of Law for providing their support and expertise. Special thanks to Professor David Ardia for his tremendous help on this piece and for his steadfast belief in me. I also thank the 2021 Cyberspace Law Seminar cohort for their helpful feedback. Lastly, I credit the editors of the Northwestern Journal of Technology and Intellectual Property for helping me take this piece through the final round. This piece is dedicated to my wife Ehmu and brother Locke. Any errors or misunderstandings are my own.
INTRODUCTION ............................................................................................................. 40
I. ALGORITHMIC CONTENT CURATION AS PLATFORM SPEECH ............................................. 47
   A. Algorithms: A Technical Primer ............................................................................. 47
   B. Personalized Recommendation Decisions ................................................................. 48
   C. Recommendation Decisions as Platform Speech ......................................................... 50
II. POWER OF RECOMMENDATION DECISIONS OVER USERS ........................................... 54
   A. Mis- and Disinformation ......................................................................................... 54
   B. Addiction and Mental Health .................................................................................. 58
   C. Privacy and Autonomy Harms ................................................................................. 61
III. USER PERSONALIZATION AS A COUNTERSPEECH SOLUTION ....................................... 63
   A. Counterspeech Doctrine ......................................................................................... 63
   B. User Personalization Tools as Counterspeech ......................................................... 64
   C. Doctrinal Implications ............................................................................................. 68
IV. PURSUING EFFECTIVE USER PERSONALIZATION ......................................................... 71
   A. Inadequacy of Predominant Personalization Tools ................................................... 72
   B. Achieving Adequate Personalization Tools ............................................................... 76
V. CONCLUSION .............................................................................................................. 85

INTRODUCTION

Having left a digital far right echo chamber, Caleb Cain started an educational YouTube channel to liberate people like himself. He wanted to help people break away from the toxic thought that once controlled him. But his resolve resulted in online ridicule and death threats for being a “traitor.” He even felt he had to buy a gun to protect himself from the online community he had once enjoyed.¹

It was not always like this for Cain. A few years back, he identified as a liberal; but after he dropped out of community college, he started watching many hours of YouTube in an aimless effort to find his path in life.² Gradually, he wandered into a far-right rabbit hole.³ Cain found content


² Id.

creators who spoke directly to his concerns as a young, white man: conspiracy theories for his sense of powerlessness; misogyny for his suspicion towards women; and mainstream media censorship for his fear of voicelessness in American society. As a result, his high school friends no longer recognized him as the empathetic and progressive person they had known only a few years ago. Cain had been radicalized.

Fortunately, for the most part, Cain found his way back. After major calls for reform, YouTube’s recommendations—a constantly updating personalization algorithm that considers many factors to render video recommendations—exposed Cain to a left-leaning community which combined edgy humor, calm reasoning, and factual support to debunk the far-right ideas he had consumed in bulk. Yet, his consumption of right-wing content still persisted. Alt-right “4Chan trolls,” to make matters worse, named him a traitor and threatened to hang him for creating his own left-leaning YouTube channel. This is when he felt he had to buy a gun to protect himself. Yet, faced with pressure from online communities and a recommendation algorithm that continued to curate far-right content for Cain,
a gun may not be able to protect him from once again venturing into another rabbit hole.\textsuperscript{13}

This story exemplifies the precipitous spread of mis- and disinformation across platforms\textsuperscript{14} and its radicalizing effects on society. Like Caleb Cain, millions of users are susceptible.\textsuperscript{15} This is because mis- and disinformation tends to exploit the emotions of its audience by supplying false but believable information that reinforces people’s misconceptions, biases, and identities.\textsuperscript{16} Once this content is hosted on a platform, the platform’s personalization algorithm drives engagement to this content by exposing users who are predicted to be receptive. These users then receive similarly indexed content\textsuperscript{17} while the problematic content reaches similarly profiled users.\textsuperscript{18} In this way, mis- and disinformation has the potential to reach millions in seconds,\textsuperscript{19} enabling a familiar story of radicalization like Caleb Cain’s to further unfold in our society.

\textsuperscript{13} Cf. Gough, supra note 3 (reporting that Caleb Cain believes anyone is susceptible to radicalization).

\textsuperscript{14} This article uses the word “platform” to refer to web services that host, moderate, curate, and present another’s content for the consumption of an end user. The more technical term is “information intermediary.” Overview: Information Intermediaries, OXFORD REFERENCE ONLINE (2021), https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100003398#:--text=Individuals%20and%20groups%20who%20obtain%20or%20sell%20the%20company%27s%20shares [https://perma.cc/MJ4K-HK8].


\textsuperscript{16} See, e.g., Alice Marwick & Rebecca Lewis, Media Manipulation and Disinformation Online, DATA & SOC’Y RSH. INST. 29 (2017) [hereinafter Marwick & Lewis, Media Manipulation] (“Far-right movements exploit young men’s rebellion and dislike of “political correctness” to spread white supremacist thought, Islamophobia, and misogyny through irony and knowledge of internet culture.”); cf. William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker & Jay J. Van Bavel, Emotion Shapes the Diffusion of Moralized Content in Social Networks; 114 PNAS 7313, 7313 (2017) (observing the phenomenon of “moral contagion” where each additional moral-emotional word provided a 20% increase in content diffusion but only between group members (e.g., liberals)). However, this strategy of content engagement is not unique to mis- and disinformation. See Michael Süßlaw, Svenja Schäfer & Stephan Winter, Selective Attention in the News Feed: An Eye-Tracking Study on the Perception and Selection of Political News Posts on Facebook, 21 NEW MEDIA & SOC’Y 168, 183 (2019). Yet, in the context of mis- and disinformation, this strategy reinforces destructive, harmful or prejudicial attitudes to drive content engagement.

\textsuperscript{17} Cf. Julia Angwin, Facebook Enabled Advertisers to Reach ’Jew Haters,’ PROPUBLICA (Sept. 14, 2017, 4:00 PM), https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters [https://perma.cc/6YPF-6G8C] (explaining how Facebook presents targeted advertisements as seamlessly as possible among similar content).


\textsuperscript{19} See Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest 135 (2017) (“[T]he more people who use a platform, the more useful that platform is to each
This precarious spread of mis- and disinformation and its radicalizing effects flourish in pivotal part because of personalization algorithms. In order to generate advertising revenue, platforms strive to keep users’ attention and increase engagement. Necessary to this pursuit are several items. First, platforms need vast amounts of individualized user data regarding anything from engagement histories to personal characteristics. Second, platforms analyze user data to predict future engagement patterns, thus identifying content engagement drivers such as a user’s interests, curiosities, and identities as well as their anxieties, biases, and suspicions. Finally, to implement the platform’s individualized predictive analyses, platforms employ algorithms that render personalized algorithmic curation decisions to deliver “relevant” content to each user. The more relevant and tailored the content is, the more likely an end user will engage, thereby driving ad-revenue, more data, new predictions, and increased personalization. For brevity, I refer to these personalized algorithmic curation decisions as “recommendation decisions.”


22 See Tim Wu, Blind Spot: The Attention Economy and the Law, 82 ANTITRUST L.J. 771, 788–89 (2019) (“The high-tech Attention Brokers like Google and Facebook have made much of their ability to very precisely target the right audiences and the right states of mind.”). See, e.g., Jack M. Balkin, Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation, 51 U.C. DAVIS 1149, 1155 (2018) (“The goal of the Algorithmic Society is practical omniscience: that is, the ability to know as much as possible about who is doing what, when, and where; and the ability to predict who will do what, when, and where.”); How Feed Works, FACEBOOK [hereinafter How Facebook, How Feed Works], https://www.facebook.com/help/1155510281178725 [https://perma.cc/LE8L-SSVB] (explaining that the posts seen on Facebook’s Feed personalization algorithm are influenced by an end user’s connections and individual engagement activity on Facebook).

23 Zeynep Tufekci, Engineering the Public: Big data, Surveillance and Computational Politics, FIRST MONDAY, July 2014, https://firstmonday.org/article/view/4901/4097 [https://perma.cc/Z644-4RSS] (“[T]he advent of big datasets that contain imprints of actual behavior and social network information — social interactions, conversations, friendship networks, history of reading and commenting on a variety of platforms — along with advances in computational techniques means that political campaigns (and indeed, advertisers, corporations and others with the access to these databases as well as technical resources) can model individual voter preferences and attributes at a high level of precision . . . .”).


25 This terminology was inspired by Professor Benjamin’s term “algorithm-based decision,” which he used to easily refer to “decisions made by protocols, algorithms, and other computations.”

26
Personalization and resulting recommendation decisions can be both a general boon and a grievous bane for users. On the one hand, recommendation decisions provide users the convenience of receiving increasingly more interesting, educational, and novel content for those who may not have the time or capacity to do so themselves. On the other hand, because of the platform’s commercial interest in driving user engagement, these decisions often expose users to content that tends to exploit their prejudices and insecurities. This article does not aim to minimize the important—perhaps essential—role recommendation decisions play in navigating the vast troves of content online and improving the user experience. Rather, it focuses on how platform recommendation decisions embody expressive decisions that amplify problematic content like mis- and disinformation by leveraging the very same engagement drivers that attract end users to such content.

Platform recommendation decisions not only facilitate the spread of mis- and disinformation. They also promote social media addiction and erode user privacy and autonomy. Like mis- and disinformation, addictive content drives engagement by exploiting an end user’s emotional response. But rather than leveraging biases and insecurities, addictive content takes advantage of the affective needs of its audience. Recommendation decisions then use that reality to drive engagement. The more exposure to content that fulfills an end user’s affective needs, the more they tend to exhibit behaviors associated with addiction, such as relapse and desensitization. In fact, such behavioral tendencies, as some authors argue, result in neurological changes

---

Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1447 n.4 (2013). Here, for ease and to draw focus to the resulting decisions of personalization algorithms, rather than their computational processes, I choose to refer to a personalization algorithm’s output as a recommendation decision. Further, one may naturally compare content personalization with content moderation, which involves sitewide decisions to demote or remove content from a platform’s service. Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1599, 1635 (2018) (“These methods can be either reactive, in which moderators passively assess content and update software only after others bring the content to their attention, or proactive, in which teams of moderators actively seek out published content for removal.”). This article is not about moderation, save the regulatory difficulties they present as an alternative solution. See, e.g., infra notes 124–127 and 191 and accompanying text.


28 See infra Section II.A.

29 See infra Section II.B.
that are essentially no different from those induced by physical substance abuse.\textsuperscript{30}

Further, privacy and autonomy harms flow directly from increased use of recommendation decisions.\textsuperscript{31} More personalization means an ever-recent and -accurate supply of content engagement data points.\textsuperscript{32} These data points facilitate a platform’s ability to predict what users want and need, sometimes before users even know it.\textsuperscript{33} Naturally this predictive ability enables platforms to influence users’ content engagement patterns, thereby eroding individual privacy and autonomy.\textsuperscript{34} Thus, privacy and autonomy harms, as well as social media addiction and related mental health harms, are functions of personalization algorithms.

Despite the harmful effects of personalization, critical legal discourse has been scarce. Prominent legal scholars who have written on the subject largely agree that algorithm-based decisions like recommendation decisions are speech under the First Amendment.\textsuperscript{35} In fact, Professor Tim Wu suggested that personalization algorithms, which he termed “automated concierges,” are not likely afforded protection unless “this program would return not simply a mechanical projection based on the user’s previous choices, but rather a true recommendation based on the opinions, and indeed the prejudices, of the programmer.”\textsuperscript{36}

This article goes further. If recommendation decisions are speech, I argue that the best solution to their harmful effects is robust end user counterspeech.\textsuperscript{37} This most naturally takes the form of end user personalization inputs.\textsuperscript{38} Such tools allow users to influence and control the
content-hosting platform’s recommendation decisions. Thus, they are the most direct form of user input into the platform’s individualized content curation decisions.

However, the predominant personalization tools are falling short as capable countermeasures in four distinct ways. First, personalization inputs provide ex post, not ex ante, control over content. Second, such control is limited in scope like a pinhole on a vast map of content, since users are able to directly block only one particular video or channel per input. Third, and relatedly, personalization inputs act as preferences, rather than prohibitions on similar types of content. Lastly, the most powerful personalization inputs are largely speaker-based, rather than content-based. Thus, in pursuit of user personalization as effective counterspeech, these shortcomings should be corrected and users should receive a more robust suite of ex ante content-filtering controls. In this way, the expressively weak users of today can effectively counter the algorithmic decisions of the expressively powerful platforms of tomorrow, thereby thwarting the spread of mis- and disinformation, as well as related harms resulting from platform personalization, such as content addiction and privacy and autonomy harms.

Under this approach, the role of platforms’ recommendation decisions in exacerbating some of the internet’s most salient harms merely reflects the longstanding and often-observed power divide between platforms as private corporations with the ability to manipulate masses of people for profit, and relatively speech-weak users in spaces where the Supreme Court has acknowledged that some of the most important speech is taking place. Recognizing personalization as a means of providing effective counterspeech in the modern digital public sphere could help shift the balance of expressive power from platforms to users, thereby providing our democracy a healthier and more sustainable public sphere.

This article seeks to establish the following prescriptive argument: in order to counter false, harmful, addictive, or otherwise problematic online content, social media platforms should provide a more robust suite of user ex ante content-filtering personalization tools. Part I provides a technical primer on algorithms and explains why recommendation decisions constitute platform speech under the First Amendment. Part II discusses how

---

39 See infra Section III.A.
40 See infra Section III.A.
41 See infra Section IV.A.
42 See Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017) (“These websites can provide perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard.”).
recommendation decisions—the platform’s own expression—exacerbate some of the internet’s most salient problems, namely the spread of mis- and disinformation, social media addiction and other mental health concerns, as well as privacy and autonomy harms. Part III establishes end user personalization inputs as a form of counterspeech and describes how personalization tools offer a potential, and rather quintessential, solution to redressing the role of recommendation decisions in spreading problematic content. Part IV identifies the problems with industry-standard end user personalization and offers several proposals that policymakers, platforms, and other stakeholders can take to advance more robust end user personalization. Part V briefly concludes.

I. ALGORITHMIC CONTENT CURATION AS PLATFORM SPEECH

A. Algorithms: A Technical Primer

In order to understand how platforms “speak” through algorithm-based decisions, it is helpful to understand algorithms at their most fundamental level. In essence, algorithms are a set of step-by-step instructions for executing a specific task.\(^\text{44}\) For example, a plain-language algorithm for obtaining a mathematical average of two numbers might be “[a]dd the numbers and take half of the result.”\(^\text{44}\) Thus, this algorithm consists of a mere two steps that manipulate numbers as inputs to reach a mathematical average of two numbers as the desired output. An algorithm may also consist of a decision support system that results in a medical diagnosis, given certain patient inputs such as hypertension and heart disease.\(^\text{45}\) Even something as simple as the steps for brushing teeth is an algorithm for achieving the desired output of clean teeth.\(^\text{46}\)

In the context of informational technology, algorithms consist of code, written by a person and implemented by a computer.\(^\text{47}\) Programmers write source code to craft algorithms which computers then compile to effectuate

\(^{43}\) JAMES GRIMMELMAN, INTERNET LAW: CASES & PROBLEMS 24 (2020); see also Benjamin, supra note 26, at 1447 n.4 (“There is no single accepted definition of ‘algorithm’. . . . Broadly speaking, an algorithm is a set of instructions designed to produce an output.”); Joshua A. Kroll, Accountable Algorithms, 165 U. PA. L. Rev. 633, 640 n.14 (2016) (“The term ‘algorithm’ is assigned disparate technical meaning in the literatures of computer science and other fields . . . .”).

\(^{44}\) GRIMMELMAN, supra note 43, at 24.


\(^{47}\) Burrell, supra note 45, at 3 (“Code writing is a necessary skill for computational implementation of algorithms . . . .”).
human goals. As Professor Stuart Minor Benjamin states, “[m]ore and more of our activity involves not merely the transmission of bits, but the transmission of bits according to algorithms and protocols created by humans and implemented by machines.” Although “bits” tends to refer to information transmitted through electronic signals, the fundamentals remain the same across contexts: algorithms use certain inputs to achieve desired outputs. Such outputs may be dubbed: algorithm-based decisions.

Machine learning algorithms complicate the picture. But although these algorithms are quite different from simple classifier algorithms, they share a basic structure in that, whether learned or not, the computer implements instructions that are designed to achieve a certain outcome. This is because machine learning algorithms involve predictive computational decisions that depend on a matrix of weights which continually optimizes itself with new inputs. In this way, machine learning algorithms “learn” through vast amounts of data, thereby eliminating the human hand in programming and increasing decision-making opacity beyond human understanding. Although the self-adjusting nature of these algorithms diminish the need for human control and increase decision-making opacity, programmers or developers can still adjust machine learning algorithms to achieve decisions consistent with new values, goals, or realizations.

B. Personalized Recommendation Decisions

In the context of web services that host, moderate, curate, and present another’s content for the consumption of end users, personalization algorithms refer to the set of instructions specifically tasked to curate—

48 Compare GRIMMELMAN, supra note 43, at 23–24 (explaining that computers must compile source code into object code in order to execute a task), with Desai & Kroll, supra note 46, at 24–26 (suggesting that “an algorithm’s correctness can only be established relative to a specification of its behavior,” which may simply be “informal or notional” if a programmer operates from a poorly defined goal in the first place).
49 Benjamin, supra note 26, at 1446.
50 Id. at 1446 n.1.
51 This nomenclature helps to delineate the algorithm as opposed to its outcome.
55 See Roose, Making of a YouTube Radical, supra note 1.
localize, customize, and channel—hosted content for individual users.\textsuperscript{56} They largely, if not exclusively, curate by relevance to individual users.\textsuperscript{57} To make these personalized recommendation decisions, platform personalization algorithms first require data profiles about individual end users.\textsuperscript{58} This data typically includes engagement patterns with previous content, including whether they liked, disliked, or empathized,\textsuperscript{59} the length of pauses,\textsuperscript{60} demographic data, including age, race, and geographic location, and more.\textsuperscript{61} Aggregate user data can also help profile similar end users. Using this data, the platform can then analyze an individual user’s past engagement patterns to reliably predict future ones.\textsuperscript{62} Based on these predictions, the platform can assess curate the content that is most likely to elicit engagement. This process is called content personalization.

Personalization algorithms accomplish a variety of interconnected but conceptually distinct purposes. First, they strive to keep the end user’s attention.\textsuperscript{63} Tailoring an end user’s content helps the end user stay engaged with the platform. Otherwise, users may become disinterested or bored, potentially leaving the platform or opting to take their attention elsewhere. Second, personalization algorithms increase advertising revenue.\textsuperscript{64} Platforms sell user attention to advertisers “for cash.”\textsuperscript{65}

\textsuperscript{56}\textsuperscript{56} Bozdag, supra note 18, at 215.

\textsuperscript{57}\textsuperscript{57} See, e.g., Chung et al., supra note 38, at 69 (“Feeds relevant to a specific user easily get lost in the streams of feeds of subject areas of little interest. Therefore, most news sites offer personalization options . . . .”). But see Christine Warner, This is Exactly How Social Media Algorithms Work Today, SKYWORD (May 3, 2018), https://www.skyword.com/contentstandard/this-is-exactly-how-social-media-algorithms-work-today/ [https://perma.cc/KQ29-JQZF] (including algorithms that filter by recency within the scope of a personalization algorithm).

\textsuperscript{58}\textsuperscript{58} Bozdag, supra note 18, at 213.


\textsuperscript{61}\textsuperscript{61} Bozdag, supra note 18, at 213.

\textsuperscript{62} See Michal Kosinski, David Stillwell & Thore Graepel, Private Traits and Attributes Are Predictable from Digital Records of Human Behavior, 110 PNAS 5802, 5802 (2013).

\textsuperscript{63}\textsuperscript{63} See, e.g., Wu, supra note 22, at 771–72 (characterizing platforms as attention brokers who attracts attention by offering something to the public); Julie E. Cohen, Law for the Platform Economy, 51 U. CAL. DAVIS 133, 146 (2017) [hereinafter Cohen, Platform Economy] (describing that platforms engage in multi-sided markets, serving users to keep their attention and advertisers who seek users’ attention).

\textsuperscript{64}\textsuperscript{64} Wu, supra note 22, at 772.

\textsuperscript{65}\textsuperscript{65} Id.
algorithms help platforms gather additional end user data.\textsuperscript{66} Platforms benefit from data on user engagement with individually curated content because the data are direct feedback on its recommendations.\textsuperscript{67} Fourth, the data derived from the recommendations allow the personalization algorithm to learn and tweak its future recommendations to provide even more “relevant” content to end users.\textsuperscript{68}

\textbf{C. Recommendation Decisions as Platform Speech}

Platforms embrace the image of neutrality. After all, platforms serve different constituencies, which requires them to “elid[e] the tensions inherent in their service: between user-generated and commercially-produced content, between cultivating community and serving up advertising, between intervening in the delivery of content and remaining neutral.”\textsuperscript{69} For example, YouTube holds itself out as an open, egalitarian service, one that aims “to give everyone a voice and show them the world.”\textsuperscript{70} Similarly, Facebook demonstrated its long-standing commitment to neutrality when it decided to simply flag content as “disputed,” rather than to remove such content altogether, as a means to curb the spread of mis- and disinformation on its service.\textsuperscript{71} This decision effectively amounted to Facebook’s continued resistance, or at least hesitation, to exercising an excessive degree of editorial control over the quality and quantity of hosted content.\textsuperscript{72}

Yet, platforms play an active role in the digital public sphere. As platforms strive to maintain neutrality in their service, it is easy to forget that they also exercise their own expressive influence through recommendation decisions. Just as a content creator engages in expression when they upload their own content, or an end user exercises their preference when they choose their next video, the platform imbues its recommendation decisions with its values and expressions. Simply put, the platform itself is one of many

\begin{flushleft}
\textsuperscript{66} Cohen, Exploring Echo-Systems, supra note 18, at 142.
\textsuperscript{67} \textit{Id}. at 141 (“Over the entire length of time a user participates on social media or media sites, the small actions and digital interactions, such as likes, comments, ratings, reads, views, and shares, are accumulated into large mathematical databases.”).
\textsuperscript{68} \textit{Id}. at 142 (“By allowing users to rate, not only does Netflix improve its data library, but allows the system to repurpose the data to predict likely choices a viewer may make on the platform.”).
\textsuperscript{70} \textit{About}, YouTube, https://www.youtube.com/about/ [https://perma.cc/623K-JQU2].
\textsuperscript{72} \textit{Id}. 
\end{flushleft}
speakers on its service. And a natural implication of this reality is that their apparent neutrality over digital expression is a façade.\textsuperscript{73}

As speakers in the digital public sphere, platforms benefit from First Amendment protection. \textit{Turner Broadcasting System v. FCC (Turner I)}\textsuperscript{74} required two and only two elements for First Amendment coverage: that cable operators choose what to air and that in doing so they seek to communicate some message.\textsuperscript{75} Thus, Professor Benjamin argued that any algorithm-based decision that seeks to communicate some substantive message to an end user who can recognize that message constitutes platform speech.\textsuperscript{76}

Under this analysis, products like Facebook Feed and Google’s search results would be protected speech.\textsuperscript{77} Although the user experience may differ, both Facebook Feed and Google’s search results are algorithmic outputs. They seek to communicate substantive information that the service provider deems relevant, popular, helpful, or valuable to a receptive end user.\textsuperscript{78}

Intuition may suggest that no machine can produce expression in a meaningful way. This hesitation would embody the idea that platforms are merely pushing their expressive input to the front end of algorithmic speech production (to the design and learning method of its algorithm) and leave the mere implementation of their expression for the machine learning algorithm on the back end.\textsuperscript{79} Under Professor Benjamin’s analysis, however, recommendation decisions would constitute platform speech because they also seek to communicate substantive information which the platform deems relevant for end users who are ready and able to understand the relevance of such content.\textsuperscript{80}

Professor Tim Wu analyzed this issue under a theory of de facto functionality.\textsuperscript{81} He explained that carriers and conduits would not receive coverage because they merely pass along unadulterated, unedited

\begin{footnotesize}
\begin{enumerate}
\item Anupam Chander & Vivek Krishnamurthy, \textit{The Myth of Platform Neutrality}, 2 GEO. L. TECH. REV. 400, 411 (2018) ("Law can also promote neutrality by conditioning immunities on passivity, which is a species of neutrality.").
\item 512 U.S. 622 (1994).
\item Benjamin, \textit{supra} note 26, at 1459–60 (quoting Turner I, 512 U.S. at 636).
\item \textit{Id.} at 1463, 1471.
\item \textit{Id.} at 1467, 1470–71.
\item \textit{Id.} at 1467–71.
\item \textit{Id.} at 1467.
\item See Aaron Smith, \textit{Many Facebook Users Don’t Understand How the Site’s News Feed Works}, PEW RSLCH. CTR. (Sept. 5, 2018) [hereinafter Smith, \textit{Many Facebook Users Don’t Understand}], https://pewrsrch.org/2NmcbiR [https://perma.cc/JR67-X6AV] (finding that “just 38% of Facebook users ages 50 and older say they have a good understanding of why certain posts are included in it, compared with 59% of users ages 18 to 29”).
\item Wu, \textit{Machine Speech}, supra note 36, at 1496.
\end{enumerate}
\end{footnotesize}
Further, tools like graphs, charts, and contracts do not receive coverage because they merely perform some task, rather than communicate some message. Algorithms properly fall within these “communication tool” categories when they primarily facilitate another’s communication or perform some task for the user. On the other hand, algorithms would be “speech products” when they fulfill a vessel function for the speaker’s ideas or convey consciously curated content.

Applying this theory, Professor Wu distinguished between two types of personalization algorithms, which he called “automated concierges”: ones that curate content based on past data, and more “intelligent” ones that “would return not simply a mechanical projection based on the user’s previous choices, but rather a true recommendation based on the opinions, and indeed the prejudices, of the programmer.” He argues that, as a communication tool, the former would likely not receive First Amendment protection because it merely reminds the user of past preferences. The latter resembles a speech product because of its infusion with the platform’s ideas. Thus, under Professor Wu’s analysis, recommendation decisions that resemble a regurgitation of past engagement patterns would not be protected, whereas decisions informed by past user data and platform relevance predictions would be.

While many prominent scholars who wrote on the subject of algorithmic speech largely agree that recommendation decisions count as platform speech, the Supreme Court has not had occasion to decide the issue on its merits. In Brown v. Entertainment Merchants Association, the Court came close, ruling that video games were protected speech. Since video games are entirely constructed from expressive algorithm-based

\[82\] Id. at 1497.
\[83\] Id.
\[84\] Id. at 1498.
\[86\] Wu, Machine Speech, supra note 36, at 1532.
\[87\] Id. at 1533.
\[88\] Id. at 1532 (“[This] is really nothing more than reminding the user what she already wants.”).
\[89\] Id. at 1533.
\[90\] The broad immunity provided by Section 230 presents a substantial barrier to sustaining actions arising from a platform’s editorial recommendation decisions. David S. Ardia, Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity under Section 230 of the Communications Decency Act, 43 LOY. L.A. L. REV. 373, 382 (2010) (“Although section 230 set a high bar for plaintiffs to overcome, more than a third of their claims survived preemption.”).
\[91\] 564 U.S. 786 (2011).
\[92\] Id. at 790 (“Like the protected books, plays, and movies that preceded them, video games communicate ideas—and even social messages—through many familiar literary devices . . . and through features distinctive to the medium . . . .”).
outputs, this lends support to the idea that recommendation decisions are protected speech. Further, *Miami Herald Publishing Co. v. Tornillo* held that a media company’s ability to exercise its “editorial control and judgment” falls squarely within First Amendment coverage. The Court indicated that choice of material, treatment of issues, and decisions regarding the size and content of the publication comprised a newspaper’s traditional editorial function. Thus, *Tornillo* favors the notion that recommendation decisions, as a form of editorial control, are protected speech.

Reasoning from these decisions, district courts have held that the First Amendment protects algorithm-based outputs like search engines results. *Jian Zhang v. Baidu.com* is the earliest case to take up the issue. The court held that the First Amendment shielded the defendant’s search engine results. The plaintiff sought “to hold Baidu liable for, and thus punish Baidu for, a conscious decision to design its search-engine algorithms to favor certain expression on core political subjects over other expression on those same political subjects.” This theory of liability is at odds with the idea that Baidu was merely a passive platform that delivered content in a neutral way. Thus, understanding that the First Amendment protects a

---


95 Id. at 258.

96 Id.

97 A distinction between different types of platforms is worth noting. News aggregators function in a similar way to print newspapers in the mass media context. Both compile stories from different authors and curate newsworthy information. Print newspapers, however, originate from a centralized speaker, with a singular editorial vision, while news aggregators make recommendations on stories authored by many different speakers with disparate editorial visions. Nonetheless, those recommendation decisions come from a central speaker with a singular editorial vision—presenting relevant content—such that an aggregator’s interest in maintaining editorial control remains compelling. However, when analyzing social media platforms, the differences become stark. Unlike stories compiled in the mass media context or even stories recommended in the news aggregator context, information on social media is not always newsworthy. Although the private nature and general lack of newsworthiness of recommendation decisions do not render the speech unworthy of protection, these factors may weaken the protection afforded to such decisions. This effect is further discussed in Section IV.B.2.


100 Id. at 439–40.

101 Id. at 440; see also Benjamin, supra note 26 at 1469–70; Volokh & Falk, supra note 85, at 898–99; James Grimmelmann, *Speech Engines*, 98 MINN. L. REV. 868, 868 (2014).
newspaper’s editorial choice over content, the Baidu court reasoned that the First Amendment also protected Baidu’s editorial bias for certain political content.\(^{102}\) Because recommendation decisions are derived more from platform analytics and not from user queries like search engine results, this reasoning would easily extend to recommendation decisions. In sum, these decisions, as well as the analysis of numerous leading scholars, support the proposition that recommendation decisions would be protected speech if the issue were adjudicated.

II. POWER OF RECOMMENDATION DECISIONS OVER USERS

We can understand recommendation decisions as exerting expressive power over users. As explained above, they have significant positive dimensions, delivering educational, novel, and entertaining content to users. At the same time, they expose users to false, addictive, and otherwise problematic content. Section II.A examines mis- and disinformation as an excellent example of the kind of problematic content that recommendation decisions expressively impose on relatively speech-weak users. Section II.B then argues that recommendation decisions play a similarly integral role in facilitating social media addiction and related mental health harms. Lastly, Section II.C explains how privacy and autonomy harms flow from recommendation decisions.

A. Mis- and Disinformation

If disinformation is generally defined as intentionally false but reasonably believable information that exploits end users’ anxieties, emotions, and prejudices to entice engagement, then misinformation is similarly defined as reasonably believable “false, mistaken, or misleading information” that also relies on those same subconscious levers to drive engagement.\(^{103}\) The falsity of the content, therefore, goes hand in hand with the very factors that tend to drive end users to such content. It is because of this symbiosis between content falsity and content engagement that personalization algorithms tend to spread mis- and disinformation through

---

\(^{102}\) See Jian Zhang, 10 F. Supp. 3d at 440 (concluding that the theory of liability—punishing Baidu’s “conscious decision to design its search-engine algorithms to favor certain expression”—would violate Baidu’s editorial autonomy).

recommendation decisions. Until a platform chooses to intervene, these algorithms autonomously render decisions that expose—and re-expose—end users to mis- and disinformation to drive engagement.

Both YouTube and Facebook have well-documented problems with mis- and disinformation. For example, videos spreading disinformation in the form of conspiracy theories run rampant on YouTube, often garnering tens of millions of views in the span of a few days. Sometimes set against an unnerving soundtrack, this type of content would “unspool[] a series of far-fetched hypotheses” like how iPhones record every word uttered in its proximity, how California wildfires were part of a purposeful insurance fraud scheme, and even how “the mass shooting at Sandy Hook Elementary School in 2012 was a hoax perpetrated by gun control advocates.”

Even without factual support, these kinds of videos drive engagement by playing on viewers’ skepticism towards official power structures and fascination with gaming the system. Based on engagement patterns and user data profiles, YouTube’s personalization algorithm tends to curate these videos for end users predicted to be receptive to conspiracy theories and the underlying emotional drivers they elicit.

In late 2019, recognizing the role of its recommendation decisions in spreading mis- and disinformation, YouTube vowed to “start[] reducing recommendations of borderline content and disinformation.”


106 ALGROTRANSPARENCY, supra note 7.

107 E.g., All Things Considered, ‘Facebook Groups Are Destroying America’: Researcher on Misinformation Spread Online, NPR (June 22, 2020), https://www.npr.org/2020/06/22/881826881/facebook-groups-are-destroying-america-researcher-on-misinformation-spread-online [https://perma.cc/BZ45-QMDB] (“[F]acebook groups ha[ve] been the primary vector of disinformation on Facebook for the past several years.”).


109 Id.

110 Id.

111 Marwick & Lewis, Media Manipulation, supra note 16, at 18. See generally Cass R. Sunstein, Deliberative Trouble? Why Groups Go to Extremes, 110 YALE L.J. 71, 112 (2000) (“[M]embers will move to positions that lack merit but are predictable consequences of the particular circumstances of enclave deliberation. In the extreme case, enclave deliberation may even put social stability at risk (for better or for worse).”).

112 See supra Section I.B.
or videos that could misinform users in harmful ways.” At the same time, YouTube effectively acknowledged that it can control its recommendation decisions not to expose users to content deemed akin to conspiracy theories and other forms of misinformation.

Spreading misinformation has many problems but two are especially important: falsity and radicalization. The first is rather obvious. The spread of false information is a harm to end users and the society they find themselves in. Truth is not monolithic. Yet, promoting the spread of falsity leads end users further away from objective reality, thereby weakening their ability to engage in society in a meaningful way—as voters, employees, and consumers. This atomization of a common reality is exactly what recommendation decisions tend to facilitate by spreading misinformation.

With falsity comes radicalization. By allaying fears, confirming suspicions, and rationalizing apparent discrepancies, misinformation has an emotional appeal which captures and converts end users into walking, talking embodiments of the content individually curated for them. Naturally, in an effort to keep users coming back for more, the platform’s individualized recommendation decisions become more extreme,

---

113 Roose, Conspiracy Theory Boom, supra note 7.
114 Id. Due to growing scrutiny around its recommendation decisions, TikTok has also made the same acknowledgment. Mia Sato, TikTok Says It’s Varying For You Recommendations To Avoid Harmful Content Holes, VERGE (Dec. 16, 2021, 10:31 AM), https://www.theverge.com/2021/12/16/22839453/tiktok-for-you-recommendations-harmful-content-fyp [https://perma.cc/KC34-P2EQ] (“TikTok is changing its For You Page algorithm to prevent users from seeing too many similar videos that could be harmful when viewed in clusters . . . .”).
117 See generally Cohen, Exploring Echo-Systems, supra note 18.
118 See supra notes 1–19 and accompanying text. The story of Caleb Cain illustrates this very point. As described in the introduction, Caleb Cain was once a young and impressionable man who sympathized with identity rights, feminism, and other liberal ideological commitments, but YouTube’s recommendation decisions fed him a consistent stream of misinformation. This content helped him make sense of the world as he saw it, becoming his truth, although much of the recommendations helped reinforce a false and radical narrative of the world.
as do end users’ preferences.\textsuperscript{119} This is necessarily an incremental change occurring through repeated exposure to relevant mis- and disinformation until users’ attitudes stray far enough to alter their previous behaviors and beliefs. This process of radicalization can lead to even poorer consequences in society, such as “weakened group ties, a lack of adherence to social norms, fragmentation of identity, and purposelessness.”\textsuperscript{120} Sometimes, it may even lead to tragic death.\textsuperscript{121}

The problem of mis- and disinformation is made worse by the fact that it is quite stubborn.\textsuperscript{122} Studies have repeatedly shown that when end users are confronted by demonstrably true but opposing viewpoints, they double down on their false beliefs.\textsuperscript{123} Thus, content-based solutions like content moderation tend to be ineffective, as well as expensive, over- and underinclusive, and legally presumed invalid as a regulatory modality.\textsuperscript{124}

Further, calling out disinformants merely emboldens such actors and amplifies the spread of mis- and disinformation. For example, disinformation campaigns “did not stop once Russia’s IRA was caught interfering in the 2016 election.”\textsuperscript{125} Instead, engagement with disinformation increased and the range of “public policy issues, national security issues, and issues pertinent to younger voters” merely broadened in response to the negative press.\textsuperscript{126} Thus, speaker-based solutions tend to exacerbate the problem. Given the difficulty of curbing the harmful effects of mis- and disinformation after

\textsuperscript{119} See Ribeiro et al., supra note 3, at 140.
\textsuperscript{120} Marwick & Lewis, Media Manipulation, supra note 16, at 29.
\textsuperscript{121} Dylann Roof’s attack on a historic black church in 2015 resulted in the deaths of nine African Americans and “was predicated by his radicalization into white nationalism through his internet habits.” See Cohen, Exploring Echo-Systems, supra note 18, at 146 (“Crowd-based judgments about relevance can create information cascades that lend sensationalized, false, and hatred-inciting online material extraordinary staying power.”).
\textsuperscript{122} Cohen, Platform Economy, supra note 63, at 149.
\textsuperscript{124} Klonick, supra note 26, at 1631–32 (describing platforms’ use of both an expensive yet vague standards-based approach and a cheap yet over- and underinclusive approach to content moderation). See generally John Samples, Why the Government Should Not Regulate Content Moderation of Social Media, CATO INST. (Apr. 9, 2019), https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media [https://perma.cc/6CMJ-S7Y3] (arguing against content moderation regulation because the First Amendment likely precludes such regulation and “tech firms appear determined to deal with such harms, leaving little for the government to do”).
\textsuperscript{126} Id.
publication and exposure, an architecture-based solution curbing exposure before it happens would be promising.\textsuperscript{127}

\textbf{B. Addiction and Mental Health}

Recommendation decisions also play a crucial role in facilitating social media addiction and related mental health harm among a variety of end users.\textsuperscript{128} The way recommendation decisions produce social media addiction is similar to the way they spread mis- and disinformation: content is curated based on the very same engagement drivers that result in behavioral addictive tendencies among end users. More specifically, personalization algorithms peddle content that is predicted to give the end user pleasure or satisfaction, which are the very same factors that keep end users coming back for more.\textsuperscript{129} Although other factors like graphical user interface optimization and differential reward interval timing also facilitate social media addiction, the content that is curated specifically to produce increasing levels of social media use is fundamental to increasing platform engagement and thus keeping users hooked.\textsuperscript{130}

This process can lead to a bona fide addiction to social media use.\textsuperscript{131} There is no firm consensus on sufficiently problematic social media use as to constitute an addiction,\textsuperscript{132} but one longstanding theoretical model has

\begin{footnotesize}
\textsuperscript{127} See infra Section IV.B.1; cf. Klonick, supra note 26, at 1603 (“The first solution to this problem should . . . [come] from simple changes to the architecture and governance systems put in place by these platforms.”).  

\textsuperscript{128} See, e.g., Rosenquist et al., supra note 30, at 444 (suggesting that personalization algorithms allow social media platforms to be more appealing to users such that “social media is the source of much of the malicious content that is hypothesized to be a key cause of mental health harm”).  

\textsuperscript{129} See id.  

\textsuperscript{130} See id. at 444, 446.  

\textsuperscript{131} Framing problematic social media use as a bona fide addiction could be said to diminish the gravity of other addictions such as those based on substance abuse. Cf. Adam Alter, Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked 3–4 (2017) (“We tend to think of addiction as something inherent in certain people—those we label addicts.”). However, social media addiction is similar to the addiction induced by gambling. Id. at 5–6 (mentioning gambling and other addictive activities such as online shopping and email). In fact, the criteria used for operationalizing social media addiction are adapted from the criteria used for diagnosing pathological gambling. See Eduardo Guedes, Federica Sancassiani, Mauro Giovanni Carta, Carlos Campos, Sergio Machado, Anna Lucia Spear King & Antonio Egidio Nardi, Internet Addiction and Excessive Social Networks Use: What About Facebook?, 12 CLINICAL PRAC. & EPIDEM. MENTAL HEALTH 43, 45 (2016) (“The Internet Addiction Scale . . . was developed by adapting DSM-IV criteria for pathological gambling, a diagnosis classified as an impulse-control disorder.”).  

\end{footnotesize}
defined social media addiction using six criteria. A social media addiction may be characterized as (1) increased preoccupation or content consumption; (2) the use of social media to reduce negative feelings; (3) the gradual tolerance to pleasure derived from consumption; (4) distress or discomfort from nonuse; (5) impairment or harm in daily life as a result of social media use; and (6) consistent failures to control social media use. According to this theoretical model, “any behavior (e.g., social networking) that fulfills the aforementioned six criteria can be operationally defined as an addiction.”

Research has found that all six of these criteria positively correlate with end users who show signs of narcissism and low self-esteem. This finding suggests that addiction may arise when recommendation decisions satisfy an end user’s all too human need for self-importance and self-affirmation towards increasing levels. Other research has substantiated the same phenomenon of social media addiction from general societal trends. Time spent on social media has been growing at an accelerating rate. Further, survey data shows that relapse is increasingly prevalent among significant populations of end users who express a desire to reduce their use. Thus, these two general trends—the exponential rise of use and the increasing prevalence of relapse—suggest that social media consumption can, and has, become addictive.

Social media’s addictive qualities also tend to disproportionately impact vulnerable populations. Young, single girls are more likely to develop problematic social media usage. Indeed, in late 2021, an internal study
conducted by Facebook (now Meta) showed that Instagram use resulted in body image issues and other mental health problems among adolescent and teenage girls.\textsuperscript{143} Further, there is a growing body of evidence that suggests that, “[w]hile the benefits seem to be generally higher among younger, better educated, and the White racial/ethnic groups, the harms seem to be higher among older, less educated, and minority racial/ethnic groups.”\textsuperscript{144}

Although recommendation decisions drive engagement with content that tends to exacerbate adolescent mental health, there is evidence that positive (active) engagement with such content, such as commenting on others posts, alleviates users’ feeling of inadequacy.\textsuperscript{145} However, when adolescent social media use is moderate to high, increases in positive use correlate with increases in loneliness over time.\textsuperscript{146} Thus, even though some active use can be beneficial to adolescents and perhaps other at-risk populations, extensive use would still result in harm to users’ well-being.\textsuperscript{147} Yet, users may find it difficult to escape such harm to their well-being when recommendation decisions foster engagement with addictive content at increasing levels.

In sum, although no solid consensus has emerged on the topic of addiction and other mental health hazards arising from social media use, the topic itself has become increasingly salient in our society. Much of this discourse around the addictive effects of social media deals with graphical user interface optimization and differential reward interval timing, but the

\textsuperscript{143} Jonathan Haidt, \textit{The Dangerous Experiment on Teen Girls}, ATLANTIC (Nov. 21, 2021), https://www.theatlantic.com/ideas/archive/2021/11/facebooks-dangerous-experiment-teen-girls/620767/ [https://perma.cc/FP4G-N77E]. In many ways, this news was worrisome but not surprising, since Instagram is the one platform where user-generated content primarily consists of pictures showing dazzling lifestyles, glittering material possessions, and heteronormative body figures. LEV MANOVICH, INSTAGRAM AND CONTEMPORARY IMAGE 111, 126 (2017) (arguing that Instagram promotes an aesthetic society where “production and presentation of beautiful images, experiences, styles, and user interaction designs is central for its economic and social functioning” and “strong rules one has to follow to attract many followers” emerges). A culture of content showcasing traditional notions of wealth and beauty will naturally exacerbate feelings of inadequacy in at-risk youth.

\textsuperscript{144} Mesfin A. Bekalu, Rachel F. McCloud & K. Viswanath, \textit{Association of Social Media Use with Social Well-Being, Positive Mental Health, and Self-Rated Health: Disentangling Routine Use from Emotional Connection To Use}, 46 HEALTH EDUC. & BEHAV. 69, 78 (2019).

\textsuperscript{145} Id. at 70–71.

\textsuperscript{146} Kexin Wang, Eline Frison, Steven Eggermont & Laura Vandenbosch, \textit{Active Public Facebook Use and Adolescents' Feelings of Loneliness: Evidence for a Curvilinear Relationship}, 67 J. ADOLESCENCE 35, 42 (2018) (explaining that this relationship may be the result of heavy active Facebook use displacing real-life social connections as well as the fact that “heavy Facebook users might feel lonelier because they are the most prone to engage in social comparison processes”).

\textsuperscript{147} Cf. id. (understanding that the U-shaped relationship between active Facebook use and social and emotional loneliness suggested the “desensitization, habituation and ceiling effects” from increasing social media use that was once effective in alleviating social and emotional loneliness).
other kind of addictive feature worth our attention is the platform’s ability to render addictive recommendation decisions for masses of end users.

C. Privacy and Autonomy Harms

The erosion of individual privacy and autonomy through technological innovations like social media has seen extensive coverage in recent years. However, the way in which recommendation decisions in particular erode these interests have not yet been the subject of robust conversation. Only when we understand the broader implications such decisions have on privacy and autonomy are we able to further contextualize the problems of mis- and disinformation and social media addiction.

As mentioned above, recommendation decisions themselves are a means of gathering data. The quality and quantity of user engagement with these decisions are a valuable source of direct feedback on the quality of its recommendations. Data gathering innovations like mouse and eye tracking, coupled with traditional notions of user engagement like watch time, likes, topics, and more, allow platforms to gain insight on whether an end user’s recommendation decisions are properly keeping their attention and driving their engagement. Not only do these curation decisions themselves offer many different kinds of data points on user engagement. They also offer significant amounts of data on users. On YouTube, recommendation decisions drive eighty-one percent of content consumption. On platforms like Facebook, Instagram, and TikTok, the percentage is higher, since the entire user experience is predicated on personalized content consumption.

---

148 See, e.g., Cohen, Platform Economy, supra note 63, at 141 (explaining the datafication of everyday life); THE SOCIAL DILEMMA (Exposure Labs & Argent Pictures 2020) (explaining how social media platforms use vast troves of data on hundreds of millions of users to optimize their systems); Will Thomas DeVries, Protecting Privacy in the Digital Age, 18 BERKELEY TECH. L.J. 283 passim (2003) (noticing an erosion of privacy and autonomy in the digital age across “the three branches of privacy law”).

149 Cohen, Exploring Echo-Systems, supra note 18, at 142.

150 Id. at 141 (“Over the entire length of time a user participates on social media or media sites, the small actions and digital interactions, such as likes, comments, ratings, reads, views, and shares, are accumulated into large mathematical databases.”).


152 Facebook’s user experience primarily centers around algorithmically curating updates from friends and family, news articles, videos, and other media through its tabs. Instagram primarily feeds its users content, first, from those the user is following, and then algorithmically curated posts from others. TikTok’s whole popularity hinges on giving users easy access to an infinite stream of recommended videos. For more information on Facebook usage trends in America, see John Gramlich, 10 Facts About Americans and Facebook, PEW RSCH. CTR. (June 1, 2021), https://www.pewresearch.org/fact-tank/2021/06/01/facts-about-americans-and-facebook/ [https://perma.cc/QEV6-TD3W].
Simply put, the more users consume content through recommendation decisions, the more data the platform can gather on its users.\textsuperscript{153} And because of the quality and quantity of the data collected, platforms are able make highly-sensitive inferences that users themselves would feel uncomfortable sharing to others, such as religious preferences and health conditions.\textsuperscript{154}

Moreover, recommendation decisions provide platforms an incentive to gather data. These decisions are only as good as the data informing them.\textsuperscript{155} Thus, platforms require more data on end users and their content preferences so that their recommendation decisions remain relevant. And if platforms’ recommendation decisions stay relevant, as they are designed to be, then platforms see greater engagement, attention, and inevitably monetary gain.\textsuperscript{156}

Thus, the successful functioning of recommendations is a powerful incentive to further erode privacy.

However, what flows from the erosion of privacy through recommendation decisions is the reduction of end user autonomy, namely violence to the end user’s beliefs and the contortion of their psyche. Recommendation decisions end up driving end users toward predicted outcomes, namely consuming problematic content, rather than being driven by end users towards users’ preferred outcomes.

In sum, recommendation decisions play different roles in different contexts. In the context of misinformation and social media addiction, they are nefarious because they use the very same engagement drivers that make the curation decisions problematic. In the context of privacy and autonomy harms, however, they act as both a means and an incentive to gather user data, driving end users away from themselves and toward the platform’s desired outcomes. Although recommendation decisions play different roles in all three contexts, if they are platform speech, then empowering users to confront such speech through architectural, legal, market-based, and normative means should alleviate all three phenomena in one fell swoop.

\textsuperscript{153} See, e.g., Emily A. Vogels, \textit{The Longer and More Often People Use Facebook, the More Ad Preferences the Site Lists About Them}, P\textsc{ew Rsch.} C\textsc{tr.} (Dec. 3, 2019), https://www.pewresearch.org/fact-tank/2019/12/03/facebook-ad-preferences-linked-to-frequency-of-use-age-of-account/ [https://perma.cc/Y26Q-R8SB].

\textsuperscript{154} Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek & Blase Ur, \textit{Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization, in CHI ’18: PROCEEDINGS OF THE 2018 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS} 1, 8 (2018).

\textsuperscript{155} See supra notes 57–62.

\textsuperscript{156} See supra notes 63–65.
III. USER PERSONALIZATION AS A COUNTERSPEECH SOLUTION

Consistent with free speech values, if recommendation decisions are platform speech, then the constitutionally preferred method for combating problematic recommendation decisions is end user counterspeech. In the digital space, this primarily occurs via user personalization tools. These tools offer end users individualized control and influence over recommendation decisions. Thus, personalization tools may be fertile regulatory ground for encouraging expressive combat from the bottom-up. Section III.A briefly explains the counterspeech doctrine. Section III.B establishes why personalization tools enable end user counterspeech against recommendation decisions.

A. Counterspeech Doctrine

The counterspeech doctrine stands for the premise that the constitutionally preferred method of overcoming harmful, injurious, false, or otherwise problematic speech is more speech (i.e., counterspeech). As Justice Brandeis famously articulated, “If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.” This is because truth, and in fact democracy itself, flourishes if information can flow freely like water and where ideas can battle for acceptance. Thus, the Supreme Court has endorsed counterspeech as the first remedy against false statements such as those that cause reputational harm. The Court has also used the doctrine to invalidate government restrictions on false speech itself, especially when the legislating state is attempting to protect its citizens from certain kinds of information.

In pursuit of this ideal, modern scholarly discourse has conceptualized counterspeech as expression that directly responds to or undermines problematic speech such as false, harmful, or injurious statements, images,

---

160 See Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) (“[T]he ultimate good desired is better reached by free trade in ideas—that the best test of truth is the power of the thought to get itself accepted in the competition of the market . . .”).
161 Richards & Calvert, supra note 157, at 554.
or other communications. Some scholars would include in this definition any speech that has incidental countereffects against problematic speech, such as public educational campaigns. Further, given the rise in problematic speech in the information era, counterspeech has seen increasing attention and popularity as a regulatory solution to extreme and dangerous speech.

Although the counterspeech doctrine rests on many assumptions, three deserve special attention. First, the doctrine assumes individuals can distinguish between truth and falsehoods. Just as consumers in an economic market can distinguish goods by quality, so too can individuals in the marketplace of ideas differentiate ideas by truth. Second, for truth to eventually prevail, individuals must “place greater value on true news and information than they do on false information.” Otherwise, adding more speech may promote the very information that counterspeech purports to extinguish. Third, the doctrine posits that counterspeech is most effective as a means towards truth only upon a wide and robust discussion of ideas. These three assumptions are worth presenting here because of their potential as sources for counterarguments against user personalization inputs as a form of counterspeech.

B. User Personalization Tools as Counterspeech

Counterspeech is great, but exactly what form does it take when countering problematic recommendation decisions? Clearly, scholars and other stakeholders engage in counterspeech when they raise an alarm

---

164 Id.
167 Id. (“[T]here is the assumption that individuals are capable of discerning between true and false information.”).
168 See id.
169 Id.
170 See Richards & Calvert, supra note 157, at 556. (“[C]ounterspeech is most effective when its proponents are able to call journalistic attention to their message, place it on the media’s agenda, and thereby exponentially increase the audience to whom the message is disseminated.”).
171 See infra Section III.C.
172 See Cohen, Platform Economy, supra note 63, at 141.
173 See supra note 7 and accompanying text.
against platforms and their invisible tendency to personalize, and privatize, the user experience. On occasion, these public efforts between platforms and stakeholders have helped create change. The other form of counterspeech is more private, direct, and frequent than public criticism and advocacy by scholars and stakeholders. It is the end user’s own personalization inputs, specifically the tools that an end user uses to control or influence their experience with a platform.

Personalization tools are the primary means by which end users can directly control or influence a platform’s recommendation decisions. These tools consist of content engagement inputs, like empathizing, commenting, or favoriting content; hiding particular posts or advertisements; following, unfollowing, or blocking specific end users; and following or subscribing to forums, end users, and topics of interest. As platforms have acknowledged, user inputs are valuable because they can help refine the platform’s recommendation decisions.

---


175 See, e.g., Community Standards, FACEBOOK, https://www.facebook.com/communitystandards/ [https://perma.cc/Y2FX-HZFHI] (assuring that the platform allows users “control over their own experience by allowing them to block, unfollow or hide people and posts” as a countermeasure against content, which acts in concert with its own moderation system against harmful, offensive, injurious, or false content that may or may not violate its Community Standards). To be clear, this is not to imply that all personalization tools confer direct control over recommendation decisions. Some do not (e.g., liking a post). This is also not to say that these tools are the only factors that control or influence recommendation decisions. They are not (e.g., post recency). However, the term “personalization tool” does suggest product features, switches, or controls that allow user input and affect a platform’s recommendation decisions for that particular end user.


177 See, e.g., Google, Manage Your Recommendations, supra note 176 (“Your activity . . . may influence your YouTube search results, recommendations on the home page, in-app notifications, and suggested videos among other places.”); What Are Recommendations on Facebook?, FACEBOOK, https://www.facebook.com/help/1257205004624246 [https://perma.cc/A8XB-DL89] (“Our goal is to make recommendations that are relevant and valuable to each person who sees them.”); How TikTok Recommends Videos #ForYou, TikTOK, https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you [https://perma.cc/6VS5-SXWU] (“Ultimately, your For You feed is powered by your
Of course, in order to count as counterspeech, an end user’s personalization inputs must represent the end user’s expression. Courts have already largely decided they do. In Bland v. Roberts, the Fourth Circuit has held that “liking” a Facebook page constitutes “pure speech” under the First Amendment. Facebook provided the Court with an explanation of what a “like” means on its platform:

The “like” button on Facebook, represented by a thumbs-up icon, is a way for Users to share information on Facebook. . . . By clicking the like button, a Facebook User generates an announcement known as a “like” story that is posted to her Profile (now Timeline) page. For example, if Jane Smith liked the UNICEF Facebook Page, the statement “Jane Smith likes UNICEF” would appear on her Profile page along with the title of the Page and an icon selected by the Page’s administrator. The Page’s title and icon function as an Internet link: another Facebook User who views the User’s Profile can click on them and be taken to the Page. If Jane Smith liked an article on CNN’s website about UNICEF’s activity in sub-Saharan Africa, the statement “Jane Smith likes this article” would appear with a link to the article.

The Court reasoned that liking a Page communicated a bona fide statement of the end user’s approval. Further, the “universally understood ‘thumbs up’ symbol” in connection to a Page conveyed the end user’s message of approval. Other courts have followed suit.

Although one may argue that Bland concerned only the “like” button, the “like” button is merely emblematic of the broader range of expressive user personalization inputs. The Bland Court found no constitutional significance between the single click of a “like” and the multiple clicks of a written statement of approval. One may also argue that the “like” button is speech only because it published a public message of the end user’s approval. Thus, personalization inputs count as speech only when communicated publicly and to other end users. However, this argument also fails. The size and purpose of an audience has no bearing on whether

feedback: the system is designed to continuously improve, correct, and learn from your own engagement with the platform . . . .”)

730 F.3d 368 (4th Cir. 2013).
Id. at 386.
Brief for Facebook, Inc. as Amicus Curiae Supporting Appellant at 5, Bland v. Roberts, 730 F.3d 368 (4th Cir. 2013) (No. 12-1671).
Bland, 730 F.3d at 386.
Id.
Bland, 730 F.3d at 386 (“That a user may use a single mouse click to produce that message that he likes the page instead of typing the same message with several individual key strokes is of no constitutional significance.”).
something counts as expression. A book is still a book if read only by the editor or by millions of readers. As a “Like” is still a like if communicated only to the platform itself or to millions of end users.\textsuperscript{185} This reasoning logically extends to other personalization inputs as well.

Platforms themselves should agree with this assessment. For instance, Facebook publicly supports counterspeech as a means to flush out racism, violence, and other extreme content—when such speech makes headlines or appears in users’ personal feeds.\textsuperscript{186} Further, Facebook suggests, in addition to the platform’s control over content moderation decisions, end users can benefit by exercising control over Facebook’s recommendation decisions through various personalization tools.\textsuperscript{187} In this way, Facebook recognizes that personalization tools provide end users expressive control where the platform cannot control, or has not bothered to control, the recommendation decisions primarily rendered in pursuit of its predictive engagement strategies.

Facebook is not alone in recognizing the value of end user personalization inputs. After TikTok faced increased scrutiny for “push[ing] young users down rabbit holes of sex and drugs, showing similar content repeatedly,” TikTok quickly announced its development of “more ways [for end users] to customize what content isn’t shown to them,” including keyword and hashtag filters.\textsuperscript{188} TikTok therefore also sees value in personalization inputs as useful counterspeech. In sum, just as a platform speaks through recommendation decisions, so does an end user through personalization tools. Their inputs help change, correct, and influence the platform’s predictions. Personalization tools, in other words, are an end user’s personal countermeasures against the expressive power of the platform’s recommendation decisions.

\textsuperscript{185} \textit{Cf., e.g.}, Benjamin, \textit{supra} note 26, at 1461 (“Communication thus seems to require, at a minimum, a speaker who seeks to transmit some substantive message or messages to a listener who can recognize that message.”).

\textsuperscript{186} \textit{Supporting the Voices That Are Engaged in Counterspeech}, FACEBOOK [hereinafter Facebook, \textit{Supporting the Voices}], https://counterspeech.fb.com/en/ [https://perma.cc/4CA9-TDT2] (“As a community, a social platform, and a gathering of the shared human experience, Facebook supports critical Counterspeech initiatives by enforcing strong content policies and working alongside local communities, policymakers, experts, and changemakers to unleash Counterspeech initiatives across the globe.”).

\textsuperscript{187} \textit{Community Standards}, FACEBOOK, https://www.facebook.com/communitystandards/ [https://perma.cc/Y2FX-HZFJ] (“People can report potentially violating content, including Pages, Groups, Profiles, individual content, and comments. We also give people control over their own experience by allowing them to block, unfollow or hide people and posts.”).

\textsuperscript{188} Sato, \textit{supra} note 114. Although TikTok is working on a feature that allows users to filter their feed by keywords and hashtags, \textit{id.,} for reasons discussed in Section IV.A, these keyword-oriented filters alone cannot adequately counter the platform’s harmful recommendation decisions.
Since personalization inputs double as an end user’s own counterspeech, a regulatory focus on enhancing those inputs would be promising. Counterspeech remains constitutionally preferred, and platforms themselves have also demonstrated support for counterspeech solutions. A regulatory focus on personalization tools not only helps enhance counterspeech capabilities, thereby bolstering user participation with the platform and content itself; it also injects more speech into the digital public sphere, unlike a regulatory focus on content removal or moderation. Thus, mis- and disinformation, as pernicious as it may be, as well as addictive content and harms to privacy and autonomy, should be dealt with affirmatively so that we might best uphold core First Amendment values. Therefore, regulating personalization tools would be low-hanging fruit as a counterspeech solution to curbing the spread of harmful online content.

C. Doctrinal Implications

Should we accept the proposition that end user inputs via personalization tools constitute end user counterspeech against a particular platform’s recommendation decisions, one may still hesitate to find personalization tools to be an effective solution, given the assumptions of the counterspeech doctrine. However, as I explain below, none of these theoretical concerns substantially weaken the viability of user personalization tools as an effective form of countering a platform’s recommendation decisions.

1. Inability to Distinguish Speech

One may argue that neither users nor platforms may be able to distinguish between true and false, harmful and innocuous, problematic or unproblematic content. Admittedly, assessing the veracity of certain types of content can be difficult. For example, deepfakes have become increasingly tricky to spot. However, end users, platforms, and researchers continue to

---

189 See supra Section III.A.
190 See, e.g., Facebook, Supporting the Voices, supra note 186. In addition to leveraging users to police their recommendation decisions, platforms may also find end user personalization tools attractive because they further distance the platform as a speaker from its recommendation decisions, thereby limiting liability for third-party content, and advance the technological motif of democratization (of news media).
191 Cf. Klonick, supra note 26, at 1603, 1666 (“[T]he biggest threat this private system of governance poses to democratic culture is the loss of a fair opportunity to participate, which is compounded by the system’s lack of direct accountability to its users.”).
192 This counterargument takes inspiration from the first assumption presented in Section III.A. See supra note 167 and accompanying text (“First, the doctrine assumes individuals can distinguish between truth and falsehoods.”).
193 See Matthew Groh, Ziv Epstein, Chaz Firestone & Rosalind Picard, Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds, 119 PNAS 1, 1 (2022). Deepfakes are videos
work together to develop new ways to detect this synthetic content with substantial accuracy.\textsuperscript{194} Although developing new strategies to counteract false content should not be taken for granted, the example of deepfakes illustrates that users, platforms, and other stakeholders do have the ability to detect increasingly complicated forms of false speech. Indeed, platforms have been able to identify, index, and demote problematic content not only based on their own intervention\textsuperscript{195} but also that of their users.\textsuperscript{196}

2. Value for Falsehood over Truth

Further, one may argue that personalization tools are not worthy of our attention because the average end user tends to value the false, harmful, and addictive more than the true, benign, and moderate.\textsuperscript{197} Thus, personalization inputs would merely allow users to filter—deliberately or subconsciously—towards the bad, thereby allowing the platform to promote the very speech their inputs should seek to extinguish as a counterspeech solution.

However, this hesitation is again misplaced for two reasons. First, users engage with content for all sorts of reasons, from entertainment to education. In fact, a vast majority of users use social media primarily for educational purposes, such as learning how to do things, buying products, and staying manufactured by neural network models to make one person appear to say something that another person said by swapping or altering their face with the original face depicted. \textit{Id.}

\textsuperscript{194} See \textit{id.} at 1, 5 (finding that “human-AI collaboration” increased accuracy from 66\% to 73\%).


\textsuperscript{196} See, e.g., Report Inappropriate Content, YOUTUBE, https://support.google.com/youtube/answer/2802027?hl=en&ref_topic=9387085 [https://perma.cc/2JGC-XDLZ] (explaining that YouTube relies on its users to report content as inappropriate so that the platform can mark that content for age restriction); Sensitive Media Policy, TWITTER, https://help.twitter.com/en/rules-and-policies/media-policy [https://perma.cc/DN2Y-LWDT] (explaining that the platform relies on end users to mark sensitive content as such so that it can place “images and videos behind an interstitial (or warning message)”).

\textsuperscript{197} Soroush Vosoughi, Deb Roy & Sinan Aral, \textit{The Spread of True and False News Online}, 356 SCIENCE 1146, 1146 (2018) (finding that false information diffused faster, farther, deeper, and more broadly than truth in all categories of information); Rosenquist et al., supra note 30, at 442 (arguing social media platforms tend to cause behavioral addictive tendencies among end users). This counterargument parallels the second assumption presented in Section III.A. See supra note 169 and accompanying text (“Second, for truth to eventually prevail, individuals must ‘place greater value on true news and information than they do on false information.’”).
informed. Thus, average engagement is far less likely to prioritize problematic content. Second, even if one believes users prioritize harmful content consumption, at least giving users an adequate ability to reject harmful content would serve society better than the alternative. This is because if users had counterspeech-level control, they could effectively reject problematic content. On the other hand, if users continue to have the same level of control as they do now, they would consume problematic content at the same levels as before—just without an opportunity to adequately control problematic recommendation decisions.

3. Private Speech as Counterspeech

Lastly, one may argue that user personalization inputs are private communications—one between the end user and the platform—whereas the traditional conception of counterspeech posits a public communication—one among stakeholders in the marketplace of ideas. Yet, there are two reasons why this argument fails. First, the expressive dynamic at issue is not a public dynamic at heart. Since the platform produces problematic private speech, the user should be able to produce corrective private counterspeech. Plus, despite the private nature of the users’ counterspeech, the user’s private inputs are effective and direct countermeasures against false, harmful, and otherwise problematic recommendation decisions. Thus, although counterspeech usually takes place publicly, it can also do so privately between the user and platform.

Second, user personalization invariably does affect the modern digital public sphere in real ways. An individual user’s personalization inputs not only influence recommendation decisions concerning other users but also


200 Therefore, with counterspeech control, a user like Caleb Cain could reject alt-right mis- and disinformation from his feed as soon as he realizes that such information is false; otherwise, he would be left with the risk of relapsing—just without adequate control. See generally supra notes 1–13 and accompanying text.

201 This counterargument parallels with the third assumption presented in Section III.A. See supra note 170 and accompanying text (“Third, the doctrine posits that counterspeech is most effective as a means towards truth only upon a wide and robust discussion of ideas.”).

202 See Benech et al., Counterspeech on Twitter, supra note 163, at 14 (explaining a “one-to-one” counterspeech exchange may be private and effective).

203 See supra Part II.

204 See Richards & Calvert, supra note 157, at 533.

205 See Cohen, Exploring Echo-Systems, supra note 18, at 141.
have the capacity to promote or demote content itself.\textsuperscript{206} Thus, although a user’s personalization inputs may be private, they also have an effect on the public.

In sum, personalization tools are analogous to a form of counterspeech in the digital public sphere. They provide end users with the ability to engage with and against the platform’s recommendation decisions. Further, although I have addressed some hesitations that stem from the inherent assumptions of the counterspeech doctrine, I recognize further discussion may be warranted. However, at bottom, as long as the counterspeech doctrine remains a pillar of First Amendment jurisprudence, personalization tools remain a viable regulatory solution to the problems our society must now confront.

IV. PURSUING EFFECTIVE USER PERSONALIZATION

There is one more objection that is unrelated to the doctrine of counterspeech but that has stood the test of time. One may argue that more powerful personalization tools could exacerbate echo chambers, leading to increased insulation to novel ideas and differing viewpoints, as well as group polarization and radicalization.\textsuperscript{207} This line of thought reaches all the way back to the 90s with the advent of the Babel Objection: “the concern that information overload will lead to fragmentation of discourse, polarization, and the loss of political community.”\textsuperscript{208}

I do not attempt to dispute this fear—as many scholars have done so already for many years.\textsuperscript{209} I do, however, believe a tailored approach can minimize the risk of exacerbating information cocooning while bolstering the benefit of end user counterspeech in the digital public sphere. To this end,

\textsuperscript{206} Cf., e.g., Smith et al., Many Turn to YouTube, supra note 27 (showing YouTube’s recommendation algorithm curating increasingly longer and more popular content).

\textsuperscript{207} See CASS. R. SUNSTEIN, REPUBLIC.COM 15, 65 (2001).

\textsuperscript{208} YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 214 (2006). Diametrically opposed to the Babel Objection is the critique that the democratizing effects of the Internet leads to concentration or consolidation in infrastructure and attention, similar to an information economy dominated by mass media. Id.

\textsuperscript{209} See Marwick, Fake News?, supra note 20, at 486 (providing a comprehensive explanation of recent scholarly coverage against the filter bubble theory of personalized search results (i.e., recommendation decisions)); William H. Dutton, Bianca Reisdorf, Elizabeth Dubois & Grant Blank, Social Shaping of the Politics of Internet Search and Networking: Moving Beyond Filter Bubbles, Echo Chambers, and Fake News 5 (Quello Ctr., Working Paper No. 2944191) (2017) (finding that end users “search for and double check problematic political information, and expose themselves to a variety of viewpoints”); Eytan Bakshy, Solomon Messing & Lada A. Adamic, Exposure to Ideologically Diverse News and Opinion on Facebook, 348 SCIENCE 1130, 1131 (2015) (“Although partisans tend to maintain relationships with like-minded contacts, . . . on average more than 20% of an individual’s Facebook friends who report an ideological affiliation are from the opposing party, leaving substantial room for exposure to opposing viewpoints.”).
Part IV presents viable proposals that would enable platforms and policymakers to address the expressive power of recommendation decisions. First, Section IV.A explains exactly how the predominant personalization tools are failing to meet the needs of end users as speakers with platforms. Section IV.B focuses on redressing specific deficiencies in those tools through platform technology, laws, market incentives, and social norms.

A. Inadequacy of Predominant Personalization Tools

Many end users have not seen, touched, or used personalization tools. A majority of American Facebook users do not understand why content appears in their Facebook Feed.\(^{210}\) Over a third believe they have no control.\(^{211}\) Over two-thirds have not even attempted to influence the content that appears on their Facebook Feed.\(^{212}\) Thus, for most end users, personalization itself remains a mystery, so personalization tools remain largely ignored. The lack of awareness around end user personalization tools and how personalization algorithms work suggests that the social norms around the predominant personalization tools today are inadequate.

Further, consider the current architecture of these tools. The prevailing set of personalization tools are flawed in four major ways. They confer end user control over a platform’s recommendation decisions that is largely: (1) ex post, (2) minimal, (3) preferential (indirect), and (4) speaker-based.

YouTube’s Home Recommendations is a concrete example which illustrates these four shortcomings.\(^{213}\) For any given video that appears on the recommendation feed, YouTube allows each end user to remove specific videos via a “Not Interested” input.\(^{214}\) In essence, this tool allows an individual end user to express that they have no interest in the platform’s recommendation decision, thus filtering out, blocking, and hiding that specific piece of content. The “Don’t Recommend Channel” input is the other way end users can directly control YouTube’s personalized video recommendations.\(^{215}\) This tool allows users to express that they do not wish to hear from a specific speaker, thus blocking content from a specific channel or content creator.\(^{216}\) Other than these, end users may avail themselves to

\(^{210}\) See Smith, Many Facebook Users Don’t Understand, supra note 80.
\(^{211}\) Id.
\(^{212}\) Id.
\(^{213}\) Although I reference YouTube specifically, my observations apply to nearly all platforms, given the substantial similarity of end user personalization inputs across platforms. See supra note 176 and accompanying text.
\(^{214}\) Google, Manage Your Recommendations, supra note 176.
\(^{215}\) Id.
\(^{216}\) Id.
liking, disliking, commenting, and watching videos in order to influence YouTube’s future recommendations.\footnote{See supra notes 59–61 and accompanying text.}

No doubt, these tools improve an end user’s experience, but upon closer inspection, several shortcomings lurk in the shadows. First, since the platform must initially recommend each video or channel, every instance of end user control requires initial exposure to curated content and content creators. Accordingly, end users must risk exposure to counteract the platform’s recommendation decisions propagating problematic content or the proponents of such content. In the context of addictive content, this reality is especially alarming, as the aforementioned dynamic is similar to requiring a quitting alcoholic to hold a drink in order for the person to reject the drink.\footnote{Cf. Rosenquist et al., supra note 30, at 439 (arguing that harmful products like “credit cards and cigarettes are useful product parallels to social media platforms”).}

Second, the effect of a user’s personalization input is minimal. Only that particular video or channel is guaranteed to be blocked as a result of each instance of user input. Indeed, Google confirms this much when instructing users on how to use its inputs.\footnote{Google, Manage Your Recommendations, supra note 176 (instructing users to tap “Not Interested” to remove specific recommended videos and “Don’t recommend channel” to “make sure that videos from specific channels don’t show up in your recommendations”).} Thus, each input constitutes a minimal, limited form of expressive control within a vast universe of billions of hours of third-party content and 31 million content creators.\footnote{YouTube by the Numbers: Stats, Demographics & Fun Facts, OMNICORE (Jan. 6, 2021), https://www.omnicoreagency.com/youtube-statistics/ [https://perma.cc/6FSQ-AZ27].} The user must therefore put in the work—a deterring amount of it—just to create an impact on the content they receive.\footnote{For a thread that indicates the typical user experience arising from the minimal power of their personalization inputs, see Pikachu1001000, REDDIT, https://www.reddit.com/r/youtube/comments/b5334l/the_not_interested_button_does_not_work/ [https://perma.cc/BYH8-WZKH].}

Third, and relatedly, most user personalization inputs effectively constitute a preference—not a prohibition—against similar types of content. Although the “Not Interested” and “Don’t Recommend Channel” inputs block specific content and content creators, the point at which those inputs change future the quality of future recommendations is often delayed “as late as possible.”\footnote{See Mitchell M. Tseng, Yue Wang & Roger J. Jiao, Mass Customization, in CIRP ENCYCLOPEDIA OF PRODUCTION ENGINEERING 7 (2017).} This strategy, formally called delayed differentiation, is widely used in industry because it improves efficiency by reducing the

\footnote{See supra notes 59–61 and accompanying text.}
\footnote{Cf. Rosenquist et al., supra note 30, at 439 (arguing that harmful products like “credit cards and cigarettes are useful product parallels to social media platforms”).}
\footnote{Google, Manage Your Recommendations, supra note 176 (instructing users to tap “Not Interested” to remove specific recommended videos and “Don’t recommend channel” to “make sure that videos from specific channels don’t show up in your recommendations”).}
\footnote{YouTube by the Numbers: Stats, Demographics & Fun Facts, OMNICORE (Jan. 6, 2021), https://www.omnicoreagency.com/youtube-statistics/ [https://perma.cc/6FSQ-AZ27].}
\footnote{For a thread that indicates the typical user experience arising from the minimal power of their personalization inputs, see Pikachu1001000, REDDIT, https://www.reddit.com/r/youtube/comments/b5334l/the_not_interested_button_does_not_work/ [https://perma.cc/BYH8-WZKH].}
\footnote{See Mitchell M. Tseng, Yue Wang & Roger J. Jiao, Mass Customization, in CIRP ENCYCLOPEDIA OF PRODUCTION ENGINEERING 7 (2017).}
number of computational steps needed to make recommendations. However, by the same token, it tends to preserve and protect the use of harmful engagement drivers throughout multiple iterations of end user personalization inputs. In other words, despite an end user’s efforts to leverage their personalization inputs to target certain kinds of content, the platform will continue to delay the decision to deinstitutionalize harmful recommendation decisions. Considering that this strategy merely provides more opportunity for the platform to confirm, and for the user to inadvertently fall prey to, the very content that causes harm like mis- and disinformation and addictive content, personalization tools effectively provide little, if any, guarantee that similar types of content or content creators will be filtered out. Like a captive audience, end users risk repeated exposure even when they realize the problematic nature of the type of content they have been consuming.

Fourth, end users must rely on speaker-based filtering in order to impose any sort of categorical control over their recommendation decisions. Besides YouTube’s “Not Interested” input, the “Don’t Recommend Channel” input is the other way users may exact direct control over the content that YouTube has curated for them. This input is speaker-based because it filters out curated content based on who provides that content, not what is provided. Although this tool allows users to impose some degree of categorical control over the recommendation decisions they receive, it risks over- and under-blocking content, since the user must choose among speakers (i.e., creators), not speech (i.e., content) itself.

Facebook’s Feed illustrates this flaw in more detail. Facebook’s Feed is a popular personalization algorithm that curates “status updates, photos, videos, links, app activity and likes from people, Pages and groups that you follow on Facebook.” Facebook instructs that in order to control this personalization algorithm, end users may prioritize, unfollow, reconnect, and snooze Pages, people, and groups. Influencing how Facebook Feed ranks

---

224 See Lee & Teng, supra note 223, at 48–49 (“One can obtain additional savings in inventory cost if one can reverse a high-valued operation at an early stage with a low value-added operation at a later stage.”).
225 See id. at 49; see also Google, Manage Your Recommendations, supra note 176 (“Your ‘Not Interested’ and ‘Don’t recommend channel’ feedback may be used to tune your recommendations.”).
226 One need only look to the story of Caleb Cain and his continued exposure to and consumption of far-right content even after he consciously decided to dig himself out of the culture.
227 Facebook, How Feed Works, supra note 23 (emphasis added).
228 Control What You See in Feed, FACEBOOK [hereinafter Facebook, Control What you See], https://www.facebook.com/help/964154640320617/?helpref=hc_fnav [https://perma.cc/F8JA-H5L8].
content is more indirect.\textsuperscript{229} The ranking considers the end user’s quality and frequency of engagement with types of posts (e.g., videos), Pages, people, and groups.\textsuperscript{230} It also considers factors that are not in the end user’s control, namely post recency and the quality and frequency of the engagement of the end user’s own social network.\textsuperscript{231}

Yet, these tools are somewhat crude. To exact direct control on Facebook, the end user must largely filter by speakers.\textsuperscript{232} According to Facebook, end users must prioritize, unfollow, reconnect, and snooze content creators to exert some influence or control over Feed.\textsuperscript{233} Thus, since Feed curates problematic content based on the activity of the end user’s own social network,\textsuperscript{234} an end user must choose to unfollow or demote a member of their own social network to exact some degree of categorical control over problematic recommendation decisions.

Additionally, just as friends on Facebook are important to the end user, groups also improve an end user’s experience, cultivating trust among users based on shared traits or interests.\textsuperscript{235} Yet, since these groups are susceptible to the spread and amplification of disinformation,\textsuperscript{236} end users would have to leave to directly control any problematic recommendation decisions arising from a group. Granted, an end user could merely ignore or demote problematic content pulled from their social network or their groups to influence the recommendations they receive; yet as explained above, the end user would risk continued exposure to harmful, injurious, or false content curated by Facebook. Plus, this control must wrestle with curation decisions that pull from a vast sea of similar content. In sum, on Facebook, the most powerful, though crude, way to exert control over recommendation decisions is through speakers, not content.

\textsuperscript{229} Facebook describes its Feed ranking as providing “a personalized and diverse stream of posts from the people, news sources, businesses and communities you’ve connected with on Facebook.” Facebook, \textit{How Feed Works}, supra note 23. Facebook states that this ranking helps end users “connect people to the posts that matter to them most.” Id.

\textsuperscript{230} Id.

\textsuperscript{231} Id.

\textsuperscript{232} Pages, people, and groups on Facebook are analogous to channels on YouTube, since they provide content which the personalization algorithm pulls to render its recommendation decisions.

\textsuperscript{233} See Facebook, \textit{How Feed Works}, supra note 23 (explaining “[w]hat influences the order of posts in my Facebook feed”).

\textsuperscript{234} See \textit{id.} (explaining “[W]hat kinds of posts will I see in Feed on Facebook?”); Facebook, \textit{Control What You See,} supra note 228 (listing the “following options to adjust your Feed preferences:” favoriting, unfollowing, reconnecting, snoozing, and reacting to content and content creators).

\textsuperscript{235} All Things Considered, supra note 107.

\textsuperscript{236} Id. (“Facebook groups are ripe targets for bad actors, for people who want to spread misleading, wrong or dangerous information.”).
Some platforms allow users to control their exposure by the broad category of sensitive or offensive content. Twitter allows users to place a warning on sensitive material before exposure. YouTube provides a Restricted Mode that “use[s] many signals—such as video title, description, metadata, Community Guidelines reviews, and age-restrictions—to identify and filter out potentially mature content.” Although these tools provide for some categorical content-based control in countering a platform’s recommendation decisions, users may tend to fear the lack of control and possible over- and under-blocking. Thus, these broad categorical controls create a chilling effect on user personalization.

**B. Achieving Adequate Personalization Tools**

The problem is clear. The predominant personalization tools available to end users (1) require exposure before an end user can exert control, (2) confer a woefully limited level of control over content-filtering, (3) situate user input as merely user preference, and (4) largely depend on speaker-based filtering for a crude kind of categorical control. What are platforms and policymakers to do about these deficiencies? Structured by four well-known regulatory modalities, Section IV.B will analyze proposals for platforms, policymakers, and scholars to consider.

1. **Architecture**

From an architectural perspective, personalization tools should address these four shortcomings. A natural solution would be to allow users to filter out content by type (e.g., conspiracy theory) or by characteristics (e.g., profanity, sexual content) before end user exposure. Platforms have already acknowledged their ability to filter out and target different types of problematic content, such as conspiracy theories, offensive materials, and even fake news and disinformants. A granular ex ante content-based filtering system would allow users to consciously block false, addictive, and other problematic content with a far greater sweep than currently available. However, more importantly, this standard of end user personalization would allow users to prevent recommendation decisions from exploiting their

---


238 Turn Restricted Mode On or Off on YouTube, YOUTUBE, https://support.google.com/youtube/answer/174084?hl=en [https://perma.cc/77S2-LSFU].

239 See supra note 114 and accompanying text.

240 See supra note 195 and accompanying text.

241 Samples, supra note 124 (reporting that Facebook is using machine learning to identify and remove fake accounts); see also Manjoo, supra note 71 (reporting that Facebook opted to mark misinformation as disputed).
receptiveness to problematic content like mis- and disinformation and addictive content.

Relatedly, platforms could make their personalization tools more prominent or striking. On YouTube, Facebook, and Twitter, the personalization tools are hidden behind three small monochrome dots in vertical or horizontal formation. Further, pressing these small dots also presents an array of tools, some of which can be unrelated to end user personalization. Despite being an industry standard, this method of presentation can be hard to notice and confusing to use, especially for the less-technologically inclined. Therefore, to raise end user awareness for personalization tools, platforms should consider implementing more intuitive or striking product designs, or de-cluttering menus where personalization tools reside. (Perhaps swiping left to reject misinformation, instead of a “Not Interested” input buried in a little menu behind three grey dots.)

Interestingly, platforms have recently implemented prominent information panels for certain types of problematic content. These panels appear when a particular piece of content touches on topics prone to mis- and disinformation, such as election fraud claims, the COVID-19 pandemic, and even the moon landing. Although these panels probably help curb the harmful effects of mis- and disinformation by providing valuable corroborating or opposing third-party viewpoints, these panels may still risk users glossing over them due to their density and unattractive design. More importantly, these panels work only after the user is already drawn into some particular piece of content. Thus, they allow recommendation decisions to succeed and leave the hard work of scrutinizing the information on the back end. Further, since mis- and disinformation evokes emotional responses and users tend to be less receptive to opposing viewpoints after exposure, these

---


244 E.g., Information Panel Giving Topical Context, GOOGLE, https://support.google.com/youtube/answer/9004474?hl=en [https://perma.cc/N7YW-Q3CV] (explaining that these panels appear regardless of “what opinions or perspectives are expressed in a video” and provide independently-sourced background information).

245 Id.
information panels may not prove effective, like slapping on an ugly Band-Aid to cure a broken bone. The architecture-based solution articulated above is superior because it focuses on combating problematic recommendation decisions themselves and allows users to counter the problematic content before exposure.

2. Law

Policymakers could consider directly requiring platforms to follow the standard of personalization tools proposed above. The academic and legislative focus has been on content moderation, so there is little direct law implicating content personalization. As a result, a regulation on personalization tools has not yet been attempted. However, such a regulation would inevitably present difficult First Amendment hurdles. In this section, although I do not attempt to resolve those issues, I argue that a direct regulatory approach could be viable upon passing intermediate scrutiny.

To start, case law on filtering regulations suggests that Congress could, and should, consider regulating and promoting personalization tools. In *Ashcroft v. ACLU*, the Supreme Court suggested that Congress should directly promote filtering technology.

One argument to the contrary is worth mentioning—the argument that filtering software is not an available alternative because Congress may not require it to be used. That argument carries little weight, because Congress undoubtedly may act to encourage the use of filters. We have held that Congress can give strong incentives to schools and libraries to use them. It could also take steps to promote their development by industry, and their use by parents.

Although the Court did not expressly certify filtering as a regulatory means, the Court clearly indicated a preference for content filtering over content removal under the First Amendment. Since filtering technology encompasses algorithms that maneuver the user through hosted content, the Court’s approval logically extends to personalization tools as well.

---

246 See, e.g., Counterspeech, DANGEROUS SPEECH PROJECT, https://dangerousspeech.org/counterspeech/ [https://perma.cc/AHV9-5NJM] (“Continued study of counterspeech is essential, especially as censorship and takedown proliferate as methods of regulating online speech.”).


248 Id. at 669.

249 Id. (emphasis added) (citations omitted).

250 In recent years, the legislature has been less approving of consumer choice over platform decisions. In 2016, the Federal Communication Commission (FCC) passed a rule that required telecommunication carriers to provide their customers with participatory tools to grant or deny the carriers’ use of their sensitive personal information for marketing and other purposes. Protecting the Privacy of Customers of Broadband and Other Telecommunications Services, 81 Fed. Reg. 87274 (Dec. 2, 2016). Yet, Congress took swift action to disapprove of and nullify the rule. See Joint Resolution, Pub. L. No.
Despite the Court’s apparent enthusiasm for filtering technology, regulations on personalization tools must still pass muster under the First Amendment. The first question in this analysis is whether the proposed regulation targets or burdens the platform’s speech.\textsuperscript{251} A regulation requiring platforms to implement the user personalization standard described above would not impose any content- or viewpoint-based restriction on speech. The proposed regulation would be generally applicable because it would merely impose content-filtering \textit{capabilities}, not content filtering itself. In other words, it will merely require additional filters that will allow users to make more choices over the content filtering they experience on the platform.

However, the proposed regulation would likely be restricting platforms’ right to free speech by placing an incidental burden on its right to produce certain recommendation decisions for end users.\textsuperscript{252} In today’s legal landscape, platforms have free reign to render any recommendation decisions for its end users. Compliance with the proposed regulatory standard could suppress vast swathes of decisions that would otherwise be available. Given this potential effect, it may even chill platforms’ recommendation production, since platforms could suppress their production in fear of over-triggering end user personalization. Thus, given the potential impact on recommendation decisions, intermediate scrutiny would probably apply such that courts would require the government to show that the regulation is “no greater than is essential” to support an important government interest.\textsuperscript{253} Because intermediate scrutiny does not necessarily spell the death of a well-tailored regulation, legislators may find a viable path toward regulating end user personalization to achieve the promise of counterspeech.

Beyond the platform’s right to render recommendation decisions, several other speech-related interests converge when analyzing the viability of a proposed regulation on end user personalization. A platform’s right to communicate and to editorialize compete with the end user’s “very basic

\textsuperscript{115} 22, 131 Stat. 88 (2017). However, the past three years have seen heightened appetite for regulating platforms such that consumer choice may seem more appealing to Congress.

\textsuperscript{251} United States v. O’Brien, 391 U.S. 367, 375, 376 (1968) (noting first “that the [challenged statute] plainly does not abridge free speech on its face”).

\textsuperscript{252} Cf. United States v. Am. Library Assoc., 539 U.S. 194 (2003) (indicating that a statute which conditions funding on public libraries’ use of filtering software interferes with the libraries’ right to editorialize their collection as well as the patron’s right to receive information).

\textsuperscript{253} See Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 661–62 (1994) (applying intermediate scrutiny for “content-neutral restrictions that impose an incidental burden on speech”). Several dimensions of recommendation decisions further justify a lower level of scrutiny than strict scrutiny, including their private nature, see supra Section I.C; commercial purpose, see Sofia Grafanaki, \textit{Platforms, the First Amendment and Online Speech: Regulating the Filters}, 39 PACE L. REV. 111, 151 (2018); and their largely mathematical or mechanical, rather than artistic or creative, production process, see Wu, \textit{Machine Speech}, supra note 81, at 1533.
right to be free from sights, sounds, and tangible matter [they] do not want” and their right to reply. Two court cases illustrate why this interest competition matters. The first is *Rowan v. U.S. Post Office Department*. In *Rowan*, the Supreme Court prioritized a statutory right to block unsolicited mail, despite its impact on advertisers’ fundamental right to communicate. The Court held that a statute giving homeowners a right to provide notice against, reject, and filter out postal solicitations did not violate the First Amendment because the privacy of the home is sacred and the regulation empowered homeowners to target speech, rather than “vesting the power to make any discretionary evaluation of the material in a governmental official.”

The Court also believed that holding otherwise would violate the homeowner’s autonomy, stating the following:

“[I]t would make hardly more sense than to say that a radio or television viewer may not twist the dial to cut off an offensive or boring communication and thus bar its entering his home.”

In *Tornillo*, the Supreme Court ruled in favor of the speaker’s autonomy rather than the audience’s right to reply. In a unanimous decision, the Court found that a state statute that conferred to political candidates a right to reply to material critical to their candidacy in local newspapers violated the First Amendment. The statute imposed a content-based penalty for material that was critical of political candidates, since compliance required newspapers to absorb additional costs in printing, composing, and publication space, thereby also producing a chilling effect on political controversy, since newspapers would probably rather avoid the material altogether than absorb the costs imposed by a right to reply. The Court declared, even if no additional costs were imposed, the statute constituted an impermissible “intrusion into the function of editors,” who are entitled to

---

254 *Compare* Miami Herald Pub. Co. v. Tornillo, 418 U.S. 241, 258 (1974) (weighing the newspaper’s right to editorialize with a politician’s right to reply to critical material), with *Rowan v. U.S. Post Office Dep’t*, 397 U.S. 728, 736 (1970) (weighing the advertiser’s right to communicate unsolicited mail with the homeowner’s right to block such mail). Although personalization tools do not allow end users to publicly reply to critical material of themselves through an intermediary, as usually contemplated, these tools do allow end users to respond to the platform’s recommendation decisions with their own edits. *Cf. Tornillo*, 418 U.S. at 258 (assessing a politician’s statutory right to reply to critical material). See generally Patrick M. Garry, *The Flip Side of the First Amendment: A Right to Filter*, 2004 Mich. St. L. Rev. 57, 69 (2004) (characterizing user filtering as user editing).


256 *Id.* at 737–38.

257 See *id.* at 737.

258 *Tornillo*, 418 U.S. at 258.

259 *Id.* at 247, 258.

260 *Id.* at 256–57.
make decisions as to its publication’s size and content, as well as its

treatment of certain public issues and officials.  

On its face, platform-user personalization seems more similar to the

expressive dynamic in Rowan than in Tornillo. Like the choice to reject

unsolicited mail in Rowan, personalization tools reflect the choice of the user
to reject the platform’s invasive recommendation decisions. Although

recommendation decisions, unlike junk mail, are an integral part to a

consented, privatized service, their largely commercial purpose as well

as their practical invisibility in operation and technical derivation renders

decisions analogous to the unsolicited, commercial junk mail at issue in Rowan.

One difference worth its weight is that Rowan was concerned with

maintaining the privacy of the home rather than the privacy of a commercial

service. Yet, social media has become an integral part of everyday life. In

fact, social media is more invasive than traditional print media and even

broadcast media for many reasons. One reason is because it remains as

accessible as “modern cell phones.” Thus, under Rowan, the asserted right

of a platform in rendering recommendation decisions, like the asserted right

of mailers in delivering unwanted mail to the home, should “stop[] at the

outer boundary of every person’s domain.” At the very least, the similarity

between the expressive dynamic in Rowan and the one at issue here should
call into question whether communicating unwanted recommendation
decisions should constitutionally trump the end user’s ability to provide

notice against, reject, and filter out such communications.

Under Tornillo, however, the platform’s editorial right in rendering

recommendation decisions could trump the user’s right to provide

personalization inputs. However, in the personalization context, Tornillo
does not fit as well as Rowan. The assumption in Tornillo is that the

newspaper’s editorial rights and the politicians’ right to reply are mutually

exclusive. Thus, even though politicians’ right to reply benefited readers


decisions as to its publication’s size and content, as well as its

treatment of certain public issues and officials.  

On its face, platform-user personalization seems more similar to the

expressive dynamic in Rowan than in Tornillo. Like the choice to reject

unsolicited mail in Rowan, personalization tools reflect the choice of the user
to reject the platform’s invasive recommendation decisions. Although

recommendation decisions, unlike junk mail, are an integral part to a

consented, privatized service, their largely commercial purpose as well

as their practical invisibility in operation and technical derivation renders

decisions analogous to the unsolicited, commercial junk mail at issue in Rowan.

One difference worth its weight is that Rowan was concerned with

maintaining the privacy of the home rather than the privacy of a commercial

service. Yet, social media has become an integral part of everyday life. In

fact, social media is more invasive than traditional print media and even

broadcast media for many reasons. One reason is because it remains as

accessible as “modern cell phones.” Thus, under Rowan, the asserted right

of a platform in rendering recommendation decisions, like the asserted right

of mailers in delivering unwanted mail to the home, should “stop[] at the

outer boundary of every person’s domain.” At the very least, the similarity

between the expressive dynamic in Rowan and the one at issue here should
call into question whether communicating unwanted recommendation
decisions should constitutionally trump the end user’s ability to provide

notice against, reject, and filter out such communications.

Under Tornillo, however, the platform’s editorial right in rendering

recommendation decisions could trump the user’s right to provide

personalization inputs. However, in the personalization context, Tornillo
does not fit as well as Rowan. The assumption in Tornillo is that the

newspaper’s editorial rights and the politicians’ right to reply are mutually

exclusive. Thus, even though politicians’ right to reply benefited readers

make decisions as to its publication’s size and content, as well as its
treatment of certain public issues and officials.  

On its face, platform-user personalization seems more similar to the

expressive dynamic in Rowan than in Tornillo. Like the choice to reject

unsolicited mail in Rowan, personalization tools reflect the choice of the user
to reject the platform’s invasive recommendation decisions. Although

recommendation decisions, unlike junk mail, are an integral part to a

consented, privatized service, their largely commercial purpose as well

as their practical invisibility in operation and technical derivation renders
decisions analogous to the unsolicited, commercial junk mail at issue in Rowan.

One difference worth its weight is that Rowan was concerned with

maintaining the privacy of the home rather than the privacy of a commercial

service. Yet, social media has become an integral part of everyday life. In

fact, social media is more invasive than traditional print media and even

broadcast media for many reasons. One reason is because it remains as

accessible as “modern cell phones.” Thus, under Rowan, the asserted right

of a platform in rendering recommendation decisions, like the asserted right

of mailers in delivering unwanted mail to the home, should “stop[] at the

outer boundary of every person’s domain.” At the very least, the similarity

between the expressive dynamic in Rowan and the one at issue here should
call into question whether communicating unwanted recommendation
decisions should constitutionally trump the end user’s ability to provide

notice against, reject, and filter out such communications.

Under Tornillo, however, the platform’s editorial right in rendering

recommendation decisions could trump the user’s right to provide

personalization inputs. However, in the personalization context, Tornillo
does not fit as well as Rowan. The assumption in Tornillo is that the

newspaper’s editorial rights and the politicians’ right to reply are mutually

exclusive. Thus, even though politicians’ right to reply benefited readers

make decisions as to its publication’s size and content, as well as its
treatment of certain public issues and officials.  

On its face, platform-user personalization seems more similar to the

expressive dynamic in Rowan than in Tornillo. Like the choice to reject

unsolicited mail in Rowan, personalization tools reflect the choice of the user
to reject the platform’s invasive recommendation decisions. Although

recommendation decisions, unlike junk mail, are an integral part to a

consented, privatized service, their largely commercial purpose as well

as their practical invisibility in operation and technical derivation renders
decisions analogous to the unsolicited, commercial junk mail at issue in Rowan.

One difference worth its weight is that Rowan was concerned with

maintaining the privacy of the home rather than the privacy of a commercial

service. Yet, social media has become an integral part of everyday life. In

fact, social media is more invasive than traditional print media and even

broadcast media for many reasons. One reason is because it remains as

accessible as “modern cell phones.” Thus, under Rowan, the asserted right

of a platform in rendering recommendation decisions, like the asserted right

of mailers in delivering unwanted mail to the home, should “stop[] at the

outer boundary of every person’s domain.” At the very least, the similarity

between the expressive dynamic in Rowan and the one at issue here should
call into question whether communicating unwanted recommendation
decisions should constitutionally trump the end user’s ability to provide

notice against, reject, and filter out such communications.

Under Tornillo, however, the platform’s editorial right in rendering

recommendation decisions could trump the user’s right to provide

personalization inputs. However, in the personalization context, Tornillo
does not fit as well as Rowan. The assumption in Tornillo is that the

newspaper’s editorial rights and the politicians’ right to reply are mutually

exclusive. Thus, even though politicians’ right to reply benefited readers


261 Id. at 258.

262 See supra note 27 and accompanying text.

263 See supra Section I.B; see also Grafanaki, supra note 253, at 151 (advocating that personalization algorithms produce commercial speech).

264 See, e.g., Cohen, Platform Economy, supra note 63, at 140 ("Additionally, the intentional ‘invisibleness’ of the algorithm further complicates how users can gain awareness of how to read their feeds and act more intentionally.").

265 Modern cells phones “are now such a pervasive and insistent part of daily life that the proverbial visitor from Mars might conclude they were an important feature of human anatomy.” Riley v. California, 573 U.S. 373, 385 (2014).

266 397 U.S. at 728.

267 Tornillo, 418 U.S. at 256–58.

81
and the public at large through robust discourse, the direct result would have been the newspaper’s lost space, profits, and journalistic autonomy over its own publication. In contrast to editorial decision in the print media context, recommendation decisions depend in large part on individual user inputs (data and personalization). Further, unlike newspapers, platforms generally fulfill their editorial function in rendering recommendation decisions by curating, not creating, a voluminous mass of content. Personalization is also a private dynamic—one between the platform and the individual end user. Whereas editing in mass media context gives the editor wide expressive impact, the same cannot be said about curating in digital media context. The foregoing distinctions emphasize that a platform’s editorial decisions depend greatly on user taste, while newspapers must allocate more weight to other considerations. Accordingly, under Tornillo, a user’s right to reply should receive more weight relative to a platform’s right to editorialize than did a politician’s right to reply relative to a newspaper’s right to editorialize. In any case, the right to editorialize may not be the death knell to a statutory right to reply in the platform-user personalization context as it was in the newspaper-politician mass media context.

In sum, a regulation targeting end user personalization is viable because intermediate scrutiny would likely apply. Further, Rowan provides an argument that an end user’s right to reject recommendation decisions trumps the platform’s right to impose them. Although Tornillo held that a newspaper’s right to editorialize trumps a politician’s right to reply, Tornillo is sufficiently distinguishable in the personalization context.

3. Markets
Amending Section 230 to provide immunity upon implementing a “counterspeech” standard of end user personalization tools may be fruitful. Section 230 of the Communications Decency Act of 1996 provides platforms immunity from civil liability based on the hosting and removal of third-party content in certain circumstances. Amending Section 230 has wide support from members from both sides of the aisle. Following this trend, one proposal could be to give platforms “Good Samaritan” immunity only if they implement a set of end user personalization tools to meet the standard identified above. For example, as recent as June 2020, the Department of Justice suggested

---

268 See supra Section I.B.
“a provision requiring an interactive computer service to have easy and apparent mechanism for users to flag unlawful content in order to benefit from Section 230 immunity. The mechanism should be reasonable based on the size and nature of the interactive computer service.”

In this way, policymakers could incentivize platforms to pursue an adequate set of personalization tools. As explained above, Ashcroft signaled approval for such a regulatory scheme under First Amendment doctrine.

To the extent that the constitutionality of conditioning immunity and conditioning funding turn on the same free speech issue, United States v. American Library Association provides a promising analog. The Supreme Court found that public libraries’ use of content filtering software as a condition for federal funding is consistent with their First Amendment rights and that of their patrons. The Children’s Internet Protection Act (“CIPA”) required public libraries to install content-filtering software that blocks access to material that would be harmful to minors in order to receive federal assistance for Internet access and computer hardware. To determine whether CIPA’s content-filtering provisions were facially constitutional, the Supreme Court had to analyze whether compliance induced public libraries to violate the First Amendment rights of their patrons to receive protected material and whether CIPA imposed an unconstitutional forfeiture of public libraries’ right to curate their collection.

As for the first issue, the Supreme Court reasoned that public libraries’ decisions to source and collect materials was not subject to heightened scrutiny because libraries had editorial discretion in deciding which content they procured. Further, given the magnitude of content on the Internet, because a library cannot make individualized editorial judgments for all the content available on their computers, “it is entirely reasonable for public libraries to reject that approach and instead exclude certain categories of content.” The Court also addressed the natural tendency for filtering technology to over- and under-block constitutionally protected speech. Though filtering technology could erroneously block permissible material while impermissible material could scrape by, the addressability of the filtering system—the ability for patrons to easily “ask a librarian to unblock

---

270 Id. at 18.
271 See supra note 248 and accompanying text.
273 Id. at 214.
274 Id. at 201.
275 See id. at 203.
276 Id. at 208.
277 Id. at 195.
[content] or (at least in the case of adults) disable the filter”—convinced the Court that the provision did not violate the speech rights of their patrons.278

As for the second issue, the Court declared that even if CIPA provided libraries with an “unconstitutional condition” claim, requiring libraries to surrender their right to curate content for their patrons in exchange for federal funding, the Constitution entitled Congress to define the limits of its federal funding programs.279 To the extent that public libraries have traditionally excluded the explicit material CIPA had proscribed and that CIPA’s filtering provision helped carry out the purpose of the funding, Congress had full authority to require filtering for federal funding.280

Conditioning some benefit, whether that be federal funding or immunity, on compliance with implementing a standard of personalization tools would implicate the same issues and conclusions. For not only are personalization tools merely a form of filtering, but also the additional element of user control over the filters—a distinctive difference between the proposal here and the statute in American Library Association—is merely a stronger version of the very dimension the Court found attractive about CIPA’s approach: users have discretion to block or unblock material. Thus, American Library Association illustrates that the First Amendment poses little threat to an indirect approach like amending Section 230 to condition immunity on compliance with a standard of filtering.

4. Social Norms

Platforms and policymakers should also consider raising awareness about end user personalization tools. Tutorials on how to use personalization tools would go a long way in establishing awareness for user control over the platform’s recommendation decisions. Although educational campaigns have proven to be a modest step towards changing consumer behavior, they can and have played an important role in establishing new social norms. Yet, rather than a broad education campaign, platforms and policymakers themselves have the ability to provide platform-specific tutorials on how end users might take advantage of the personalization tools at their disposal.281

A platform-specific approach would likely prove more useful and effective

278 Id. at 209.
279 Id. at 213.
280 Id.
281 One example is Google’s explanation for how its search algorithm works. How Search Works, GOOGLE, https://www.google.com/search/howsearchworks/ [https://perma.cc/AH8P-ASAC]. However, one must click through and stop at the “kids & family” feature to find any discussion on content filtering and supervision controls. Helpful Features, GOOGLE, https://www.google.com/search/howsearchworks/features/ [https://perma.cc/MES2-T9M8]. Google may be able to establish more of a norm around end user personalization if it expanded the audience for such information and made such information easier to find.
than a broad campaign towards raising public awareness for personalization tools, since users would benefit from an explanation of which tools are available on a platform, what they do, and where they can find them.

More broadly, decreasing user’s susceptibility to recommendation decisions through media literacy can help mitigate the effects of the algorithm-based spread of problematic content like mis- and disinformation and addictive content. Researchers have called for educators to teach skepticism when consuming content. This approach can help lower instances of problematic recommendation decisions because end users would be able to not only recognize decisions that exploit their emotions but also alter their own engagement patterns so as to gradually inform the personalization algorithm towards different—healthier—curation decisions.

V. CONCLUSION

Why should policymakers bother with regulating towards more powerful end user personalization tools when platforms are trending towards more sophisticated personalization systems anyways? Policymakers could just do nothing. However, platforms constitute some of the most important places for speech, and personalization algorithms are the primary means with which users find information. Yet, since these algorithms are a function of the practical, commercial purposes of the platforms, they use the very same drivers that make problematic content so compelling. Therefore, the increasing spread of mis- and disinformation and addictive content is directly related to the increasing roles personalization algorithms and recommendation decisions play in today’s information society.

Thus, personalization tools help users engage in expressive combat against recommendation decisions spreading problematic content. Recognizing end user personalization as a means of providing effective counterspeech against such influential yet often-overlooked platform speech will help shift the balance of expressive power from platforms to users. With largely ex post, contained, preferential (indirect), and speaker-based control over a platform’s recommendation decisions, end users currently lack a robust ex ante content-filtering system that would empower users to participate and avoid content like mis- and disinformation or addictive content. Although the current personalization tools are a great start, the current system lacks such basics that it cannot achieve the promise of counterspeech, the most natural and speech-protective way to combat

problematic speech. Improving personalization tools—that is, improving participatory capabilities—will give our democracy a healthier and more sustainable digital public sphere.