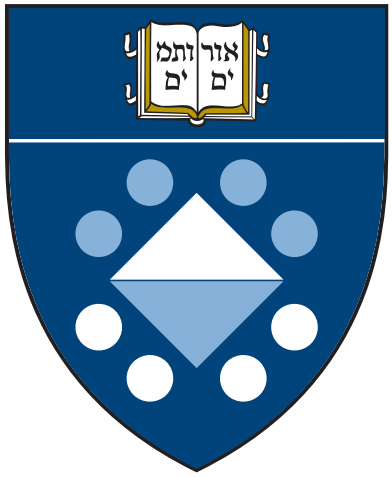

Yale University

Thurman Arnold Project

Digital Platform Regulation Conference

Paper 4

Regulating Content Recommendation Algorithms in Social Media

August 2022

Sachin Holdheim

In October of 2020, Frances Haugen captured political and national attention when she testified before the Senate Commerce Committee's Sub-Committee on Consumer Protection, Product Safety, and Data Security. She spoke of the harm created by Facebook's relentless pursuit of profits that she witnessed first-hand as a product manager.[1] In particular, she honed in on a certain kind of algorithm that she felt drove some of the most pernicious and destructive effects on society: content recommendation algorithms (CRAs).[2]

CRA is a catchall term that refers to the set of algorithms that play a part in determining the order of content displayed to an end user. These algorithms are not only used by social media companies. For example, they are also used by Google, to determine the order to list relevant search results;[3] Netflix, to display content that the company believes will be appealing to the user;[4] and Amazon, to facilitate consumer purchases.[5] In the social media context, though, these algorithms take on outsize importance: CRAs are so crucial to the operation of social media companies that they are referred to not just as one *type* of algorithm, but "the" algorithm and are often personified in discussion (i.e., "the TikTok algorithm knows me so well!").[6]

Prior to Haugen's testimony, much of the discourse around algorithmic harm focused on the ways that algorithms could directly perpetuate existing inequalities in the physical world. For example, early literature focused on the discriminatory effects of algorithmic bail setting, hiring, or policing.[7] Yet content recommendation algorithms foster a different type of harm: one that distorts community perceptions, discourse, and values in harmful, polarizing, or violent ways.[8] New forms of regulation are needed to address these urgent concerns.

This paper builds off of Haugen's conceptions of algorithmic harm in social media content recommendation algorithms to propose three classes of policy solutions. Part 1 traces the

---

[1] Frances Haugen, Statement at the United States Senate Committee on Commerce, Science, and Transportation Sub-Committee on Consumer Protection, Product Safety, and Data Security (Oct. 4, 2021) (transcript available at https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49).

[2] *Id.* ("Right now, Facebook chooses what information billions of people see, shaping their perception of reality. Even those who don't use Facebook are impacted by the radicalization of people who do. A company with control over our deepest thoughts, feelings and behaviors needs real oversight.").

[3] *How Search Algorithms Work*, GOOGLE SEARCH (last visited May 20, 2022), https://www.google.com/search/howsearchworks/algorithms/.

[4] *How Netflix's Recommendations System Works*, NETFLIX HELP CENTER (last visited May 20, 2022), https://help.netflix.com/en/node/100639.

[5] Larry Hardesty, *The History of Amazon's Recommendation Algorithm*, AMAZON SCIENCE (Nov. 22, 2019), https://www.amazon.science/the-history-of-amazons-recommendation-algorithm.

[6] *See, e.g.*, Wall Street Journal, *How TikTok's Algorithm Figures You Out*, YOUTUBE (July 21, 2021), https://www.youtube.com/watch?v=nfczi2cI6Cs; Ben Smith, "How TikTok Reads Your Mind", N.Y. TIMES, Dec. 5, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html/.

[7] *See generally* Cathy O'Neil, WEAPONS OF MATH DESTRUCTION (2017) (analyzing the use of everyday algorithmic decision-making).

[8] Haugen, *supra* note 1 ("The result has been a system that amplifies division, extremism, and polarization — and undermining societies around the world. In some cases, this dangerous online talk has led to actual violence that harms and even kills people. In other cases, their profit optimizing machine is generating self-harm and self-hate — especially for vulnerable groups, like teenage girls. These problems have been confirmed repeatedly by Facebook's own internal research.").

development of CRAs and will elaborate upon new conceptions of algorithmic harm linked to these algorithms. Part 2 proposes specific policy solutions tied to three broad categories of reform: the regulation of algorithmic design, the regulation of algorithmic intent, and the regulation of algorithmic effect.

I. The History and Harm of Content Recommendation Algorithms

The rise of social media transformed society. What was once a limited way to keep in touch with close friends has spiraled into a mammoth industry with billions of users worldwide.[9] As social media has grown, the way that content is displayed on social media platforms has changed: the sheer amount of content on each platform has given rise to CRA curation.

In 2006, Facebook pioneered the News Feed: a feed that pooled content from all of a user's Facebook friends in one place.[10] The same year, Twitter debuted to the public.[11] Both platforms ordered content in reverse chronological order; content was displayed on a *literal* timeline.[12] Yet as social media use increased, the need for curated content increased as well. It became impossible for the average consumer to read all posts from all friends in the time a user spent online.[13] Social media platforms thus began to roll out content recommendation algorithms that would reorder content to promote a business goal—generally, some form of user engagement.[14] These CRAs are separate from other algorithms that may affect what content is able to be displayed on the platform: all social media platforms filter for illicit content and spam.[15] While CRA content curation does not limit what could be *posted* to the platform, it limits what would be *seen* by the average user.

Social media CRAs are constantly under development. In 2009, Facebook implemented its first curation algorithm by sorting posts by the number of "Likes" in addition to time of posting.[16] In 2016, Facebook began to order posts according to the total amount of time that

---

[9] *Number of Social Network Users Worldwide from 2017 to 2025*, STATISTA (last visited May 20. 2022). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
[10] Mark Zuckerberg, *Happy 10th Birthday, News Feed!*, FACEBOOK (Sept. 5, 2016), https://www.facebook.com/zuck/posts/10103084921703971.
[11] *This Day in History | July 15: Twitter Launches*, HISTORY (June 28, 2019), https://www.history.com/this-day-in-history/twitter-launches.
[12] *Explained: The Algorithms that Run Facebook*, ECON. TIMES (last updated Oct. 26, 2021) https://economictimes.indiatimes.com/tech/trendspotting/explained-the-algorithms-that-run-facebook/articleshow/87282136.cms?from=mdr (Facebook); Mjahr, *Never Miss Important Tweets from People You Follow*, TWITTER BLOG (Feb. 10, 2016), https://blog.twitter.com/official/en_us/a/2016/never-miss-important-tweets-from-people-you-follow.html (Twitter).
[13] *See, e.g.*, Adam Mosseri, *Shedding More Light on How Instagram Works*, INSTAGRAM (June 8, 2021), *https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works* ("But as more people joined and more was shared, it became impossible for most people to see everything, let alone all the posts they cared about.").
[14] *See infra* Section II.B.
[15] Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC, Mar. 8, 2017, https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.
[16] *Explained*, *supra* note 12.

users spent interacting with a post, even if the user did not Like or Share the post.[17] In early 2018, Facebook pivoted from its prior user engagement standard to one that promoted "meaningful social interactions."[18] Twitter, in 2016, moved from a pure reverse-chronology ordering to one that prioritized "the Tweets you're most likely to care about,"[19] based on "accounts you interact with frequently, Tweets you engage with, and more."[20] Instagram used a reverse chronology ordering in 2010, but adopted a user-engagement-based CRA in 2016 after finding that "people were missing 70% of all their posts in Feed, including almost half of posts from their close connections."[21] Because social media platforms earn revenue by selling advertisements, the more time that a user spends on the platform, the more revenue the companies generate. Content recommendation algorithms that are designed to maximize user engagement thus directly contribute to the bottom line of social media giants. As has become clear in recent years, though, this intense focus on user engagement can cause a host of personal and societal harms.

Algorithms are powerful, and when not monitored carefully, can promulgate unexpected and far-reaching harms. For example, FTC Commissioner Rebecca Kelly Slaughter found that flawed algorithmic decision-making can facilitate proxy discrimination, enable surveillance capitalism, and inhibit competition across the marketplace.[22] The Future of Privacy Forum found that automated decision-making can lead to (1) losses of opportunity, such as in employment or housing discrimination; (2) economic loss, such as in credit discrimination; (3) social detriment, through the creation of network bubbles and stereotype reinforcement; and (4) loss of liberty, through biased surveillance and even incarceration.[23]

When discussing content recommendation algorithms specifically, no one has diagnosed algorithmic harm more clearly than whistleblower Frances Haugen. Haugen found, during her time at Facebook, that when Facebook encountered conflicts between societal benefit and corporate profit, it consistently chose the latter.[24] She discovered that Facebook's dogged pursuit of user engagement has led to creeping polarization, widespread misinformation, self-harm among vulnerable populations, and even violence and death.[25]

A number of complaints that Haugen filed with the SEC lay out her diagnoses of the harm caused by Facebook's content curation regime. For example, she discloses that Facebook

---

[17] *Id.*
[18] Mark Zuckerberg, FACEBOOK (Jan. 11, 2018), https://www.facebook.com/zuck/posts/10104413015393571.
[19] Mjahr, *supra* note 12.
[20] Rumman Chowdhury & Luca Belli, *Examining Algorithmic Amplification of Political Content on Twitter*, TWITTER: BLOG (Oct. 21, 2021) https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent.
[21] Mosseri, *supra* note 13 ("In Feed, the five interactions we look at most closely are how likely you are to spend a few seconds on a post, comment on it, like it, save it, and tap on the profile photo. The more likely you are to take an action, and the more heavily we weigh that action, the higher up you'll see the post.").
[22] Rebecca Kelly Slaughter, Janice Kopec, & Mohamad Batal, *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission*, 23 YALE J.L. & TECH. 37 (2021).
[23] *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making*, FUTURE OF PRIVACY FORUM (Dec. 11 2017) https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making.
[24] Haugen, *supra* note 1.
[25] *Id.*

was aware of the use of its platform by the Burmese military government to spread fake news to support its genocide of the Rohingya people.[26] Haugen also released internal Facebook records that show that the company "only take[s] action against approximately 2% of the hate speech on the platform," despite its public statements to the contrary.[27] She disclosed internal Facebook research that found that "13.5% of teen girls on Instagram say that the platform makes thoughts of 'Suicide and Self Injury' worse" and that "17% of teen girl Instagram users say the platform makes 'Eating Issues' (e.g. anorexia and bulimia) worse."[28] Importantly, Haugen found that these concerns were not remedied by Facebook's shift to the "Meaningful Social Interaction" standard for content curation. Instead, internal records show that this standard championed "divisive and sensationalist content," and that "the more negative comments a piece of content instigates, the higher the likelihood for the link to get more traffic."[29]

The current content curation system at Facebook is causing harm to individuals, to American society, and to the world at large. Importantly, these harms are not restricted to just Facebook and Instagram. Brookings has found, based on a meta-review of available literature, that "platforms like Facebook, YouTube, and Twitter likely are not the root causes of political polarization, but they do exacerbate it."[30] A bipartisan group of state attorneys general announced an investigation into the potential harms of TikTok, a platform that is comparatively understudied.[31] So long as social media remains supported by advertising revenue, the profit motive of social media platforms will be to increase user engagement. Without strong guardrails, algorithmic harm will continue to go unchecked.

---

[26] *Global Division and Ethnic Violence: Anonymous Whistleblower Disclosure Re: Supplemental Disclosure of Securities Law Violations by Facebook, Inc.*, WHISTLEBLOWER AID (available at Keith Zubrow, Maria Gavrilovic, & Alex Ortiz, *Whistleblower's SEC Complaint: Facebook Knew Platform Was Used to "Promote Human Trafficking and Domestic Servitude"*, CBS NEWS: 60 MINUTES OVERTIME, Oct. 4, 2021, https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/).

[27] *Facebook's Removal of Hate Speech: Anonymous Whistleblower Disclosure Re: Supplemental Disclosure of Securities Law Violations by Facebook, Inc.*, WHISTLEBLOWER AID (available at Keith Zubrow, Maria Gavrilovic, & Alex Ortiz, *Whistleblower's SEC Complaint: Facebook Knew Platform Was Used to "Promote Human Trafficking and Domestic Servitude"*, CBS NEWS: 60 MINUTES OVERTIME, Oct. 4, 2021, https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/).

[28] *Teen and Mental Health: Anonymous Whistleblower Disclosure Re: Supplemental Disclosure of Securities Law Violations by Facebook, Inc.*, WHISTLEBLOWER AID (available at Keith Zubrow, Maria Gavrilovic, & Alex Ortiz, *Whistleblower's SEC Complaint: Facebook Knew Platform Was Used to "Promote Human Trafficking and Domestic Servitude"*, CBS NEWS: 60 MINUTES OVERTIME, Oct. 4, 2021, https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/).

[29] *Facebook's Algorithms and the Promotion of Misinformation and Hate Speech: Anonymous Whistleblower Disclosure Re: Supplemental Disclosure of Securities Law Violations by Facebook, Inc.*, WHISTLEBLOWER AID (available at Keith Zubrow, Maria Gavrilovic, & Alex Ortiz, *Whistleblower's SEC Complaint: Facebook Knew Platform Was Used to "Promote Human Trafficking and Domestic Servitude"*, CBS NEWS: 60 MINUTES OVERTIME, Oct. 4, 2021, https://www.cbsnews.com/news/facebook-whistleblower-sec-complaint-60-minutes-2021-10-04/).

[30] Paul Barrett, Justin Hendrix, & Grant Sims, *How Tech Platforms Fuel U.S. Political Polarization and What Government Can Do About It*, BROOKINGS, Sept. 27, 2021, https://www.brookings.edu/blog/techtank/2021/09/27/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/.

[31] Cecilia Kang, *A Coalition of State Attorneys General Opens an Investigation into TikTok*, N.Y. TIMES, Mar. 2, 2022, https://www.nytimes.com/2022/03/02/technology/tiktok-states-investigation.html.

II. Reforming Content Recommendation Algorithms

There is a clear and pressing need to address the algorithmic harm that stems from CRAs. Frances Haugen has, among other proposed reforms, suggested that a reversion to reverse-chronology content curation may stem the algorithmic harm associated with CRA curation.[32] After her testimony, some platforms, including Twitter and Instagram, made reverse chronological ordering more easily accessible to users.[33] Facebook has long offered reverse-chronology ordering as an option, but does not allow for this to be set as the default—a user must manually toggle the option on each time they access the site, severely limiting the feature's realistic use.[34] But mandatory reverse chronological ordering may be too heavy a hammer. Abandoning algorithmic content curation entirely would result in information overload to consumers, decreasing the quality of social media platforms, while potentially resulting in a lowered bottom line for the platforms—a lose-lose. Instead, there are a number of reforms to CRAs that can and should be made to significantly reduce algorithmic harm while maintaining some form of algorithmic curation.

Because of the widespread and diffuse nature of social media company operations, Congress is best suited to impose an effective regulatory regime upon CRAs used in social media. This paper recommends three distinct suites of potential regulatory changes, each of which would apply to a different stage of the algorithmic life cycle.

First, regulations aimed at algorithmic ***development*** would encourage government regulators to monitor the actual creation of and updating of CRAs *before* they are deployed. These regulations could additionally mandate the creation of an algorithmic task force internal to the company that brings together siloed teams to produce periodic reports to government monitors.

Second, regulations aimed at algorithmic ***intent*** would effectively address misinformation. The government could work with social media companies to determine what values ought to be maximized by any CRA (for example, slowing the spread of discriminatory content or content that promotes addiction, misinformation, or disinformation). This reform would ideally be coupled with an explicit agency enforcement mechanism to ensure adherence to those norms.

Third, regulations aimed at algorithmic ***effect*** would be consumer facing, focused on increasing the transparency of algorithmic decision-making to the end user. By democratically

---

[32] Kari Paul, *Facebook Whistleblower Hearing: Frances Haugen Calls for More Regulation of Tech Giant –As it Happened*, GUARDIAN, Oct. 5, 2021, https://www.theguardian.com/technology/live/2021/oct/05/facebook-hearing-whistleblower-frances-haugen-testifies-us-senate-latest-news.

[33] Jack Morse, *Twitter Decides Against Force-Feeding Users its Algorithm After All*, MASHABLE, Mar. 15, 2022, https://mashable.com/article/twitter-reverses-course-swipe-timeline (Twitter), Samantha Murphy Kelly, *Instagram Brings Back Option for Reverse Chronological Feed*, CNN BUSINESS, Mar. 23, 2022, https://www.cnn.com/2022/03/23/tech/instagram-chronological-order/index.html (Instagram).

[34] Isobel Asher Hamilton, *How to Switch your Facebook Feed to a Chronological Timeline*, BUSINESS INSIDER, Jan. 9, 2022, https://www.businessinsider.com/facebook-social-media-switch-feed-chronological-timeline-2021-11.

distributing the research technique of "sock-puppet auditing," algorithmic effect regulation would help public researchers and interested consumers to understand the way that CRAs shape content and potentially empower consumers to craft a healthier or more equitable content mix.

A. Regulation of Algorithmic Development

Regulation of algorithmic development is designed to affect an algorithm before its public deployment. The first proposed reform would enable government experts to oversee pre-deployment algorithmic testing. The second proposed reform would require social media companies that employ CRAs to develop internal algorithmic task force teams that unite previously siloed stakeholders in CRA development and report out CRA updates to government regulators.

First, the results from pre-deployment algorithmic testing should be opened up to government regulators. Even with unbiased input data and a robust normative intent, algorithms can produce undesirable and unanticipated outcomes if not tested adequately.[35] Adequate pre-deployment testing is essential to catching these negative outcomes. For example, Seattle Times technology reporter Matt Day highlighted that in 2016, LinkedIn's CRA would routinely suggest male profiles when a female name with a similar spelling was entered into the search box ("Stephanie" would return "Stephen"), but that the reverse never occurred when the hundred most common male names were searched.[36] As LinkedIn is a site commonly used for job hiring, it is not a stretch to believe that this behavior could have resulted in material harm for a class of female LinkedIn users.

The current pre-deployment testing regime for CRAs is a black box. In the aforementioned example, LinkedIn's VP of Engineering described that the company had not tested for gender bias in its CRA because it did not track gender, in a misguided attempt to *prevent* gender bias.[37] He admitted that in hindsight, it was "obvious" that removing gender from the CRA development would "blind" algorithms to gender as a potential source of bias.[38] Allowing government insight into algorithmic testing will permit testing by data scientists who are driven by harm-reduction, rather than the profit motive. Catching potential sources of bias *before* algorithms are deployed will limit the negative effects that these algorithms are able to have on society.

Second, social media companies should be required to develop an internal algorithmic task force that brings together internal stakeholders to the CRA. This task force should report out all updates to the CRA and the effects of these updates on consumers to the government on a quarterly basis.

---

[35] Kelly Slaughter et al., *supra* note 22, at 14-15.
[36] *Id.* at 18 (citing Matt Day, *How LinkedIn's Search Engine May Reflect a Gender Bias*, SEATTLE TIMES (Aug. 31, 2016)).
[37] Igor Perisic, *Making Hard Choices: The Quest for Ethics in Machine Learning*, LINKEDIN: ENGINEERING (Nov. 23, 2016) https://engineering.linkedin.com/blog/2016/11/making-hard-choices--the-quest-for-ethics-in-machine-learning.
[38] *Id.*

CRAs are not developed by a single team. The process that determines the final order of content on a consumer's screen is the product of hundreds, or even thousands of algorithms.[39] There is often little coordination between the teams that create and oversee each of the constituent algorithms; often, the teams have competing objectives.[40] Frances Haugen has stated that these competing objectives were perennially decided in favor of that which promoted greater user engagement, and thus, supported Facebook's bottom line.[41] Moreover, teams are frequently unaware of the ramifications that their own algorithm will have on the larger CRA pre-deployment. According to Krishna Gade, the former Engineering Manager for News Feed from 2016-2018, the status quo for CRA updates was simply to deploy the changed algorithm to small groups of users and monitor for reduced engagement levels before applying to the rest of Facebook's user base.[42] This experimentalist approach fails Facebook's users: the CRA's development teams must be aware of the ramifications of an algorithmic update pre-deployment, even if the test is only applied to a small slice of the user base.

Bringing together disparate and siloed algorithmic development teams into a single task force may encourage joint conflict resolution in a way that can better support a harm-reduction framework. At the very least, it could help Facebook employees to better understand, and thus better control, the competing ways in which underlying CRAs work together to form "the algorithm," helping to put the CRA back in the control of employees. Government oversight over updates to the CRA, as achieved through the implementation of a quarterly reporting requirement, would result in greater transparency to the government and thus increase regulatory ability. Even absent any action by government regulators, developer knowledge of government eyes on algorithmic updates would likely foster greater care and deliberation when altering the CRA.

Although these reforms would slow down the rapid pace of iterative CRA development and redevelopment, increased oversight—by the government and by internal developers—of the CRA development process could encourage a paradigm shift for social media companies, emphasizing a more deliberative harm-reduction based approach to CRA management and putting the brakes on the experimentalist "move fast and break things" culture that dominates Silicon Valley.[43]

---

[39] Karen Hao, *The Facebook Whistleblower Says its Algorithms are Dangerous. Here's Why*, MIT TECH. REV., Oct. 5, 2021, https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.
[40] *Id.*
[41] Haugen, *supra* note 1.
[42] Hao, *supra* note 39.
[43] Isobel Asher Hamilton, *Mark Zuckerberg's New Values for Meta Show He Still Hasn't Truly Let Go of 'Move Fast and Break Things'*, BUS. INSIDER, Feb. 16, 2022, https://www.businessinsider.com/meta-mark-zuckerberg-new-values-move-fast-and-break-things-2022-2.

B. Regulation of Algorithmic Intent

Not all CRAs are designed to carry out the same normative goals. Regulations that affect algorithmic intent would see the government work with platforms to develop a normative framework with which CRAs would be required to comply. In particular, adding a "truthfulness" standard could dramatically alter the way in which content is ranked across the social media ecosystem.

Most social media platforms publish the normative goal maximized by the company's CRA. Facebook attempts to maximize "Meaningful Social Interactions" (MSI).[44] TikTok's CRA maximizes (per the company's own translations from Mandarin) "user value," "long-term user value," "creator value," and "platform value."[45] Twitter appears to maximize for user engagement: the company has stated that its algorithmically-ordered timeline features tweets "you are likely to care about most, [chosen] based on accounts you interact with frequently, Tweets you engage with, and much more."[46] None of these companies explicitly instruct their CRAs to amplify content based on its *truthfulness* (or conversely, none attempt to explicitly diminish the distribution of false information). Repeated accusations of widespread amplification of misinformation and disinformation have led social media platforms to adopt a variety of ex-post fixes for misinformation: adding banners on content, taking down posts, demoting untrue or misleading posts in the CRA, blocking accounts that serially post misinformation, and more.[47] Yet these fixes take down content *after* the false posts have been seen by millions—as a direct result of the initial amplification that these posts receive from the company's CRA.

If the government required that "truthfulness" be explicitly considered by a CRA, known mis- or disinformation could be required to be unpromoted by any CRA. Although this would not make a company responsible for hosting misinformation on the platform, which is protected under Section 230 of the Communications Decency Act,[48] repeated violations of this new standard would ideally impose liability on companies that systematically *promote* known mis- or disinformation via their CRAs. Platforms already have the technology to flag problematic false

---

[44] *See supra* text accompanying note 18.
[45] Ben Smith, *How TikTok Reads Your Mind*, N.Y. TIMES, Dec. 5, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html.
[46] *About Your Home Timeline on Twitter*, TWITTER HELP CENTER, https://help.twitter.com/en/using-twitter/twitter-timeline. *See also* Chowdhury & Belli, *supra* note 20.
[47] *See, e.g.*, The Associated Press, *Twitter Aims to Crack Down on Misinformation, Including Misleading Posts About Ukraine*, NPR, May 19, 2022, https://www.npr.org/2022/05/19/1100100329/twitter-misinformation-policy-ukraine (discussing Twitter's responses to misinformation about the Russian invasion of Ukraine); *About Fact-Checking on Facebook*, META BUSINESS HELP CENTER (last visited May 20, 2022), https://www.facebook.com/business/help/2593586717571940?id=673052479947730 (discussing fact checking procedures on Facebook and Instagram).
[48] 47 U.S.C. § 230. A narrow carve out of Section 230 for content recommendation algorithms was proposed by Frances Haugen. *See* Roddy Lindsay, Guest Essay, *I Designed Algorithms at Facebook. Here's How to Regulate Them.*, N.Y. TIMES, Oct. 6, 2021, https://www.nytimes.com/2021/10/06/opinion/facebook-whistleblower-section-230.html ("As Ms. Haugen testified, 'If we reformed 230 to make Facebook responsible for the consequences of their international ranking decisions, I think they would get rid of engagement-based ranking.").

content[49]: this reform would shift the burden on companies to ex-ante enforcement, rather than ex-post.

This regulation should be coupled with an explicit agency enforcement mechanism to ensure that the regulation carries weight. There is longstanding precedent of agency involvement in specific industries that are prone to consumer deception and misinformation: the FTC's Funeral Industry Practices Rule (the "Funeral Rule") is a narrow, industry-specific rule that enforces a set bundle of consumer rights.[50] It exists to prevent funeral industry operators from taking advantage of particularly vulnerable consumers.[51] Industry-specific statutes or agency rulemaking should be encouraged to protect social media users who may be similarly vulnerable in the face of rampant misinformation.

C. Regulation of Algorithmic Effect

Regulations of algorithmic effects are designed to be consumer-facing and post-CRA-deployment. Although all CRAs rely heavily on inputs from consumer behavior, the process is a black box. Consumers are never told *why* or *how* specific behavior results in specific content recommendation. Moreover, consumers are fundamentally *passive* participants in the content recommendation process: consumers are given extremely limited options to attempt to directly affect the content shown by the CRA.[52] This paper proposes two reforms to remedy these problems; both will increase consumer agency through improvements in algorithmic transparency.

First, social media platforms should be required to make public the top five reasons that each piece of content is delivered to a consumer. This information should be prominently accessible on each piece of content through a clear drop-down menu choice or button. This information should be detailed; for example, if the top reason that a piece of content is recommended is because the user has interacted with similar posts in the past, there should be a link to the similar posts in question. Providing consumers with this information would dramatically increase CRA transparency. Importantly, this transparency will lay bare problematic or inappropriate connections that the CRA may make; for example, the Mozilla Foundation

---

[49] *See, e.g.*, The Associated Press, *supra* note 47 (discussing Twitter's responses to misinformation about the Russian invasion of Ukraine); *About Fact-Checking on Facebook*, *supra* note 47 (discussing fact checking procedures on Facebook and Instagram).

[50] *The FTC Funeral Rule*, FTC Consumer Advice, https://consumer.ftc.gov/articles/ftc-funeral-rule.

[51] Cindy Skrzycki, *FTC Reviews 'Funeral Rule' to Protect Bereaved*, WASH. POST, May 14, 1999, https://www.washingtonpost.com/archive/business/1999/05/14/ftc-reviews-funeral-rule-to-protect-bereaved/186598d1-97bf-4580-8b35-0b4ad235f5f3/.

[52] On many social media platforms, the only way to directly affect the CRA except through normal platform usage is through either hiding posts from a particular user or flagging an individual piece of content as something that the user does not wish to see, which hopefully will result in the user seeing less of that "type" of content. *See, e.g., Limiting Unwanted Content*, TIKTOK NEWSROOM (Apr. 24, 2019) https://newsroom.tiktok.com/en-us/limiting-unwanted-content (TikTok); *How to Control Your Twitter Experience*, TWITTER HELP CENTER (last visited May 20, 2022), https://help.twitter.com/en/safety-and-security/control-your-twitter-experience (Twitter); *What Can I Do If I See a Post I Don't Like in Instagram Search & Explore?*, INSTAGRAM HELP CENTER (last visited May 20, 2022), https://help.instagram.com/1105548539497125 (Instagram).

found in one study that 12.2% of the videos recommended by YouTube's CRA violated YouTube's own community guidelines.[53] One 10-year-old girl, who began using YouTube to watch dance videos, was recommended videos on extreme dieting that affected her health.[54] This transparency would also allow consumers to consciously change the way that they interact with a social media platform to encourage desired content recommendations, permitting them a more active role in the content recommendation process.

Second, social media platforms should offer "sandbox" content recommendation modes whereby consumers would be able to see their recommended content ordering as if they were users with different characteristics. This would be a platform in-housing of the algorithmic testing technique known as *sock-puppet auditing*, whereby false users ("sock puppets") are created to see how an algorithm reacts to the characteristics of the sock-puppet users.[55] Users in this sandbox would be able to see how their own content base would be reorganized if they altered their political party affiliation, age, or gender, for example. While these identity alterations would only exist in a sandbox, and the user's content feed would remain unchanged, this reform would dramatically increase transparency of the algorithm and make it possible for users to better identify how CRAs shape perspective and on what presumptive identity characteristics those perspectives are chosen. This would be a powerful tool for researchers to analyze covert discrimination or the spread of misinformation across gender, class, race, and party lines.

These reforms would dramatically alter the relationship between the consumer and the content that the CRA serves. Both would also be invaluable resources for researchers or agencies that monitor CRAs for unfairness, deception, or discrimination. Empowering consumers to take a front seat in their content recommendation process may not fix the harm that unchecked CRAs can impose—but it allows individual consumers to work with government regulators and academic researchers to shed light on the black box of content recommendation algorithms.

III. Conclusion

Content recommendation algorithms are designed to maximize user engagement at the expense of all else—leading to polarization, misinformation, self-harm, and violence. Lawmakers must design appropriate guardrails to minimize this harm. Through appropriate regulation of algorithmic design, algorithmic intent, and algorithmic effect, content curation can be modified to ensure a high-quality, non-harmful product for consumers.

---

[53] *YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm,* MOZILLA FOUNDATION (July 2021), https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.
[54] *Id.*
[55] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, & Cedric Langbort, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, presented to *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* at the 64th Annual Meeting of the International Communication Association (May 22, 2014) http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.