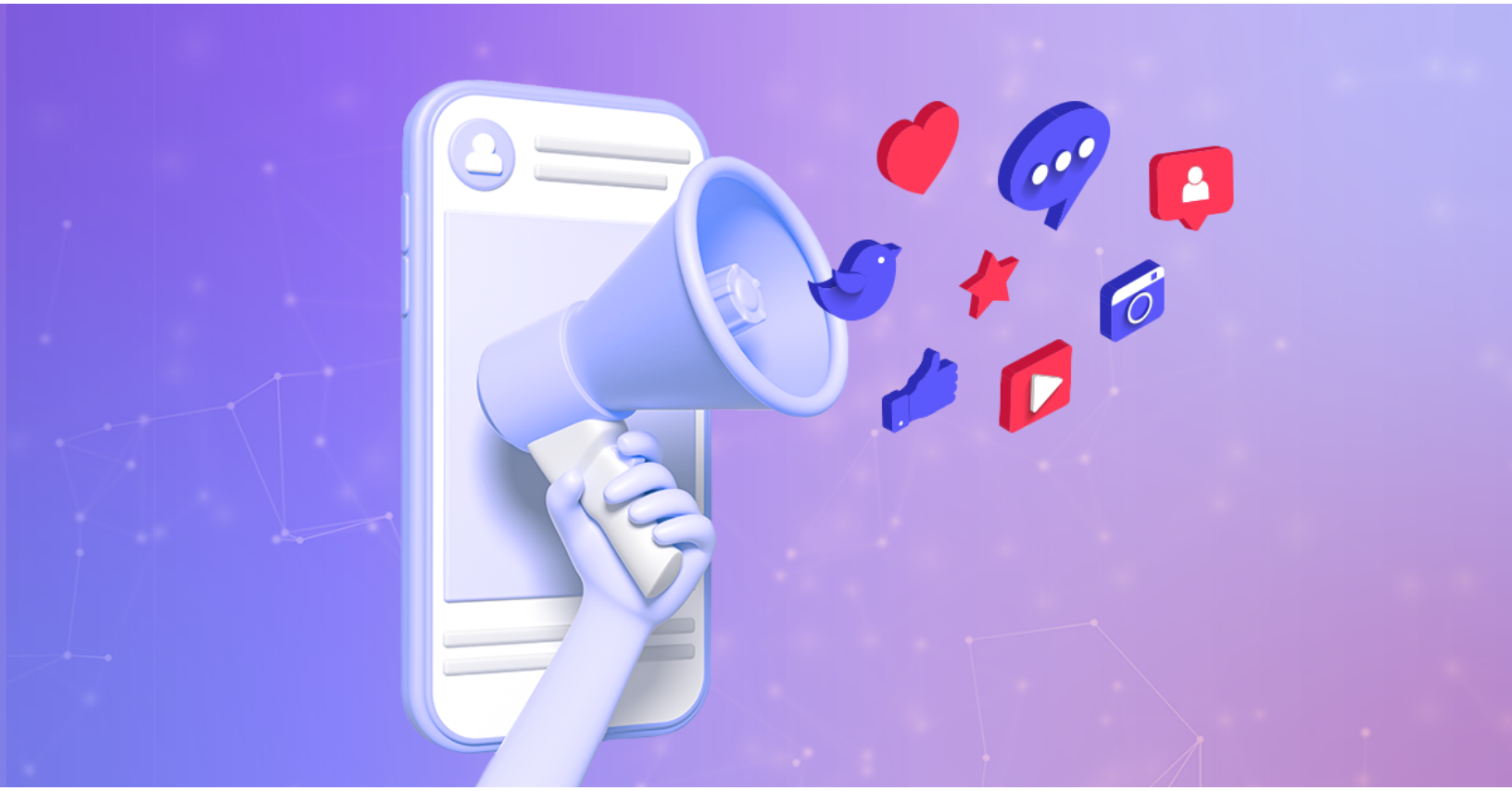
Towards Responsible Recommending

Recommendations for Policymakers & Large Online Platforms



Maximilian Gahntz

December 2022

v1.0

About Mozilla

Mozilla's mission is to ensure the internet is a global public resource, open and accessible to all. An internet that truly puts people first, where individuals can shape their own experience and are empowered, safe and independent.

Founded as a community open source project in 1998, Mozilla currently consists of two organizations: the non-profit Mozilla Foundation, which leads our movement building work; and its wholly owned subsidiary, the Mozilla Corporation, which leads our market-based work, including the development of the Firefox web browser. The two organizations work in close concert with each other and a global community of tens of thousands of volunteers under the single banner: Mozilla.

Acknowledgements

Special thanks are due to Brandi Geurkink and Owen Bennett, who helped conceive this project and offered generous support and input throughout. We would also like to thank everyone else who kindly provided feedback and advice, including: Stefan Baack; Ashley Boyd; Jason Chuang; Marshall Erwin; Michael Feldman; Karolina Iwańska; Paddy Leerssen; Raegan MacDonald; Jesse McCrosky; Kasia Odrozek; Carolyn O'Hara; Claire Pershan; Bogdana Rakova; Becca Ricks; Mika Shah; Spandana Singh; and Jonathan Stray.

Most importantly, this work draws inspiration from a growing body of work from researchers and policy professionals across academia, civil society, industry, and the public sector. We would like to thank them for their contribution to advancing the debate in this field.

Table of contents

Executive Summary

Recommender systems are a core component of many large online platforms. They significantly shape people's online (and even offline) experiences in highly automated ways, be it on social media, video streaming, or dating apps. But they are also drivers of online harms.

Much of the debate around holding platforms accountable for this focuses on mitigating negative outcomes — such as disinformation or hate — rather than the root causes, including the recommendation engines facilitating or reinforcing these outcomes. At the same time, these systems remain opaque and removed from public scrutiny. Public-interest researchers studying them struggle to gain access to high-quality data and have even been threatened by platforms. And users often lack the information and the means to effectively shape their experience on a platform.

This paper seeks to contribute to the debate by shifting focus to recommender systems and proposing a comprehensive set of actions for platforms and regulators that together can form the basis of a more responsible and accountable recommendation ecosystem, with a focus on the largest and most impactful online platforms promoting user-generated content.

In light of this, we make the following recommendations to large platforms and policymakers:

Layered oversight and scrutiny:

- Disclose **detailed information on how recommender systems work** and how they are operated, made accessible on a designated information page.
- Provide **detailed information about content demotion (or reduction) policies**, including criteria for when content is demoted and descriptions of how such policies are implemented.
- Report **aggregate data on demoted content** as part of existing transparency reporting formats.
- Provide data and tools to analyze the **most widely viewed and engaging content, pages, creators, and links.**
- Provide qualified independent researchers conducting research in the public interest with **access to data and documentation** while adhering to data protection rules and principles and robust security protocols.

- Create a **safe harbor for researchers, civil society organizations, and journalists conducting public interest research** in compliance with research ethics and data protection rules and principles.
- Conduct or commission **audits of recommender systems** and publish audit reports.

Informed and empowered users:

- Allow users to **better customize the recommendations or content displayed** to them, by providing them with more and more effective controls.
- Enable users to **exert better control over which of their data, including personal data, is collected and subsequently used** to inform recommendations.
- Allow users to **opt out of personalized recommendations** or to influence the degree of personalization at any point in their product experience.
- Provide users with **easily accessible and meaningful explanations** of why specific content is displayed to them.

These recommendations are not meant to be seen as standalone quick fixes, but rather as complementary pieces that work together to tackle key problems at their systemic roots.

Introduction

In November, voters in the U.S. took to the polls for the midterm elections. In October, Brazilians went to the booth to elect their president. In both cases, election disinformation and hate against candidates again spread widely across social media platforms both in the lead-up to and directly around the elections.[1] Similar dynamics could be observed around elections in Kenya and the Philippines earlier this year.[2] But these are only more recent manifestations of problems we have encountered again and again over the past years: Online platforms continue to contribute to harms to people and society at large. They function as important transmitters or amplifiers[3] of hate, toxicity, violence, and disinformation, often even when content is in violation of their own policies.[4] And they can be used to undermine the integrity of democratic and civic processes and debate.[5] While these are very different types of harms, they are connected by a common factor: the recommendation algorithms spreading them. Mozilla has been calling attention to this repeatedly in the past.[6]

Yet the debate around how to improve upon this status quo has often been fragmented and short-sighted. Too often it has focused on quick fixes to complex systems. Further,

---

[1] Brown and Canineu, "Social Media Platforms Are Failing Brazil's Voters"; Jeantet, "Brazilian Voters Bombarded with Misinformation before Vote"; Klepper, "As 2022 Midterms Approach, Disinformation on Social Media Platforms Continues"; Martiny, Jones, and Cooper, "Election Disinformation Thrives Following Social Media Platforms' Shift to Short-Form Video Content"; Stanley-Becker and Harwell, "Misinformation Floods the Midterms, at Times Urging Violence."

[2] Eusebio, "[ANALYSIS] Fake News and Internet Propaganda, and the Philippine Elections"; Madung, "From Dance App to Political Mercenary"; Madung, "Opaque and Overstretched, Part II: How Platforms Failed to Curb Misinformation during the Kenyan 2022 Election"; "Filipino Voters Were Engulfed in Relentless Stream of Disinformation."

[3] Although frequently used and similarly interpreted by many, the concept of "amplification" is fuzzy. Concrete definitions are sparse and varied. Keller provides a useful discussion of the difficulties associated with the term in "Amplification and Its Discontents." For the purpose of this paper, we will use Keller's definition as our working definition, i.e., understanding amplification as increasing "people's exposure to certain content beyond that created by the platform's basic hosting or transmission features." Alternatively, algorithmically amplified content might also be understood to mean content that is distributed beyond the organic reach (i.e., to subscribers, followers, friends etc.) of its creator in an automated way.

[4] For several more recent examples, see, for example, Brandt et al., "Winning the Web"; "The Facebook Files"; Integrity Institute, "Widely Viewed Content Dashboard"; Little and Richards, "TikTok's Algorithm Leads Users from Transphobic Videos to Far-Right Rabbit Holes"; McCrosky and Geurkink, "YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm"; Richards, "Examining White Supremacist and Militant Accelerationism Trends on TikTok"; The Virality Project, "Memes, Magnets and Microchips"; Thomas and Balint, "Algorithms as a Weapon Against Women."

[5] For examples, see Mozilla's recent research on the role of social media media around elections by Bösch and Ricks, "Broken Promises"; Madung, "Exporting Disinformation"; Madung, "From Dance App to Political Mercenary"; Madung, "Opaque and Overstretched, Part II: How Platforms Failed to Curb Misinformation during the Kenyan 2022 Election."

[6] "Facebook: Stop Group Recommendations"; "Tell Twitter to Pause Trends until US Election Results Are Certified."

the debate tends to focus on content *moderation* rather than *recommendation*. That is, much effort has been invested in tackling the negative outcomes — such as disinformation or hate — and how they are moderated rather than the root causes — including *how* and *why* recommendation engines distribute such content at large scale.

These recommender systems are a key and consequential product feature of many of the largest online services, helping organize vast amounts of information. By filtering, ranking, and selecting content, they determine what we see pop up on our social media feeds or video recommendations, what positions we're shown on job sites, and who we might match with on dating apps. They steer people's social, professional, and financial lives. They shape civic discourse and politics. And in many cases, they do all of this at enormous scale and in a largely automated way, with little and mostly reactive human intervention in the curation and distribution of content.[7] Yet, despite recommender systems' influence on individual and collective experiences, the major online platforms mostly fail to acknowledge their responsibility and fail to change course when business interests and the public interest are in conflict. At the same time, regulators lag behind in considering recommender systems. The EU's Digital Services Act (DSA), adopted earlier this year, is a step in the right direction in this regard with rules specifically addressing recommender systems. But we need more far-reaching change in industry and governance. The objective of this paper is therefore to set out a more comprehensive vision of what a better recommending ecosystem could look like.

While many of the recommendations put forward below may apply to recommender systems more generally, they specifically target the main distribution and ranking algorithms of the largest platforms with the highest reach and systemic impact. Further, we focus on platforms recommending user-generated content rather than "curated" platforms (e.g., YouTube rather than Netflix). We chose this focus because scale matters in this context. Platforms like Facebook, YouTube, Twitter, or TikTok have proven to have an outsized effect on public debate and how people engage with one another as well as with content online. Therefore, the stakes with respect to whether and how potentially harmful content spreads on and across these platforms increase with their size and significance.

The recommendations we offer in this paper are designed to be relevant to various actors: They are meant to point policymakers to potential pathways of regulatory intervention, particularly with regard to the largest online platforms. At the same time, our recommendations may point to models of good practice for platforms themselves. In many cases, this paper does not aim to define whether actions should be taken

[7] For an accessible overview of how such systems work, see, for example, Singh, "Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds"; Thorburn, Bengani, and Stray, "How Platform Recommenders Work."

voluntarily or mandated. That question requires dedicated consideration and will vary across contexts. This paper rather aims to set out the conditions of what a healthier recommender ecosystem would look like. In doing so, we prioritize pathways that are systemic in nature, tackling problems at their root, over quick and easy fixes.

Below, we outline the steps we think are necessary to move towards more responsible recommending. They fall under two broader aims: Ensuring layered oversight and scrutiny of platforms' recommender systems and empowering and informing users interacting with these systems.

Building a more responsible recommending ecosystem

A. Layered oversight and scrutiny

In charting the path towards a more responsible recommending ecosystem, we first must acknowledge that we are still operating under significant uncertainty: In short, we don't know what we can't know. So far, only the platforms themselves know how they're recommender systems function and how they impact users and society. They alone are in a position to gain a robust understanding of what the causal mechanisms underlying many of the problems associated with recommender systems are and how best to mitigate them. These same platforms have been reluctant to share critical information with the public, researchers and policymakers, with few exceptions. We seek but have no reliable insight into critical questions such as: How exactly and to which degree do recommender systems amplify harmful or illegal content? What is their contribution to social and political polarization in our societies? What biases are built into these systems? And how can they be abused and manipulated by malicious actors?

As publicly accessible information and robust empirical evidence on these and other questions is oftentimes lacking, we all too often are poking around in the dark and rely on makeshift approaches to finding answers. The following recommendations are therefore concerned with rectifying just that by creating systemic transparency, improving the conditions for independent research into recommender systems' design and operation, and building a layered system of oversight and scrutiny.

Disclose publicly what gets recommended — and why

While we have a high-level understanding of how state-of-the-art recommender systems function, meaningful insights into how the content presented to us online is curated are scarce and hard to find. Recommender systems are composed of a cascade of different machine learning models and rules used to source and filter prospective

content, rank it, and ultimately assemble a slate of recommendations. A deeper understanding of a recommender system therefore requires information on its components and overall architecture — yet we know little about what its constituent machine learning models optimize for[8] and the rules based on which content is filtered out and prioritized. To remedy this, public-facing documentation can be a key enabler of transparency.[9]

Companies should therefore publicly disclose detailed information on how their recommender systems work and how they are operated. This should comprise information both on the technical design and inputs of the system as well as on rules and policies affecting recommendations (see Box 1 for more detail). Additionally, this information should be as complete as possible instead of providing only select, potentially cherry-picked, examples, for instance of signals used to inform recommendations.

Box 1

Publicly disclosed information about recommender systems

Platforms should disclose publicly, amongst other things, the following information about how their recommender system functions:

- An explanation of the general architecture, logic, and design of the system
- A description of optimization goals and metrics used for the recommender system and its constituent machine learning models
- An extensive list and descriptions of the most important signals used to make recommendations (e.g. observed and inferred user data (including personal data), usage data, engagement with specific content, popularity of content)
- If, and if yes, how and based on which criteria certain types (e.g., "authoritative" content on vaccines) or sources of content (e.g., trusted news sources) are prioritized over others
- How platforms use recommender systems for enforcing content policies and guidelines, including with regard to demoted, or "borderline", content (see below)
- How users can influence or control what recommendations are displayed to them

[8] While it is conventional wisdom that platform recommender systems commonly optimize for engagement as a proxy for users' preferences, it might be too simplistic in this context as engagement can be operationalized in myriad ways.
[9] Raji and Yang, "ABOUT ML."

Currently, the little information platforms provide in this regard is often difficult to find — buried on company websites or scattered across occasional blog posts.[10] Instead, all **this information should be easily accessible on a designated information page**. Additionally, changes to the information provided on the page, like company policy changes, should be traceable over time via a change log or version history. Further, where relevant, this information should also cover regional or country-specific differences in how recommender systems are operated.

The information provided on platforms' rules and policies related to recommendations **should also cover so-called demotion or reduction policies** — often also referred to as de-amplification, shadowbanning, or as relating to "borderline" content.[11] Such policies are aimed at curbing the distribution of certain content which does not violate (but may come close) platforms' terms of service or policies. By creating and implementing such policies, many platforms have created an opaque gray area for content that they deem both acceptable and undesirable: within the boundaries of free expression and not harmful enough for removal under platforms' policies, but too harmful to be distributed widely.[12] Yet, in contrast to platforms' rules for content removal or account suspension, there is little information about how this determination is made and how such policies are implemented.

There may be good reasons for platforms to demote certain content that negatively affects users' experience or the quality of discourse on the platform (for example with regard to clickbait or misleading content) or content that is likely to be harmful but still awaits a removal decision. At the same time, such practices can materially affect users' ability to reach others and exclude them from public conversations while leaving them unaware or merely suspecting of platforms' intervention. Such policies are likely to be implemented using the same (or similar) machine learning classifiers used for flagging and automatically removing content that violates platforms' policies. Yet these tools have well-established limitations[13] and evidence suggests that wrongful removals disproportionately affect historically marginalized people, for example Black individuals discussing racial justice or LGBTQI+ people (for example, under companies' nudity policies).[14] Moreover, self-reported data from users suggests a similar pattern when it comes to content demotion.[15]

---

[10] See, for example, Goodrow, "On YouTube's Recommendation System"; Mosseri, "Shedding More Light on How Instagram Works"; TikTok, "How TikTok Recommends Videos #ForYou."
[11] For a discussion of some of these terms and their limitations, see Gillespie, "Reduction / Borderline Content / Shadowbanning."
[12] Heldt, "Borderline Speech."
[13] Shenkman, Thakur, and Llansó, "Do You See What I See?"
[14] See, for example, Haimson et al., "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users."
[15] Nicholas, "Shedding Light on Shadowbanning."

To rectify this transparency deficit, platforms should, where applicable, **provide clear information about and criteria for content they demote** (including definitions of categories like "borderline" content) and **provide a description of *how* such policies are implemented** at a technical level (e.g., is such content merely shown to fewer people or removed from recommendations or search results altogether?). Facebook's Content Distribution Guidelines[16] released in late 2021 are a step in this direction, outlining high-level categories of demoted content. However, they still often lack specificity and shed little light on the implementation of the guidelines. At the same time, we are lacking an understanding of the rate of such interventions from platforms. While content removals are reported in platforms' transparency reports, the incidence and spread of demoted content are indeterminate. As part of existing reporting structures, **platforms should therefore disclose aggregate data on the content they demote** (see Box 2).

Box 2

Publicly disclosed information about demoted content

Platforms should report the following data as part of existing transparency or enforcement reports, where possible broken down by category of content, country, and/or language:

- Aggregate data on the total amount and relative share of demoted content compared to the total volume of content on the service
- Aggregate data on how often such content has been recommended, viewed, and interacted with
- Aggregate data on how much of such content was later removed from the platform, for what reasons, how long after being uploaded, and how often it was recommended, viewed, and interacted with

Finally, the personalized nature of platforms' feeds makes it difficult to discern overall trends. Which post spreads most widely? Which creator gets the most engagement? Generally, we don't know. After all, such trends emerge from the aggregate of each user's unique recommendations. To give the public a bird's-eye view of such trends, **platforms should publish information[17] on and provide tools[18] to analyze the content, pages, creators, and links that garner the most views and engagement** on

[16] Stepanov, "Sharing Our Content Distribution Guidelines."
[17] See, for example, Meta, "Widely Viewed Content Report: What People See on Facebook."
[18] An illustrative example is the case of Meta-owned tool CrowdTangle, see Alba, "Meta Pulls Support for Tool Used to Keep Misinformation in Check"; Roose, "Inside Facebook's Data Wars."

their service. In the long run, this information might help discern what types of content recommendation systems are particularly disposed to raising to the (public) surface.

Enable more and better research on recommender systems

While there already is important research about online platforms and their recommender systems, it has so far been constrained by a lack of access to data for researchers. Instead, they have to rely on their own tools and methods for data collection, each with their own limitations. As platforms have shielded themselves and their data from public scrutiny, we still know far too little about their inner workings and their impact on users and society at large. Additionally, the research that's available has disproportionately focused on platforms providing better conditions for research because of this.[19]

While some platforms have recently moved to open up to researchers[20] — presumably also in response to upcoming data access obligations under the EU's Digital Services Act (DSA) — the overall conditions for independent research on platforms and their recommender systems remain insufficient.

Therefore, **qualified independent researchers conducting research in the public interest should gain access to non-personal and anonymized or pseudonymized data and documentation and be enabled to audit recommender systems**. This type of access should also extend to qualified researchers without academic affiliations, such as investigative journalists or researchers working with civil society organizations, as well as to third-party auditors conducting legally mandated audits (as foreseen by, for example, the DSA) or otherwise working with platforms.[21] Researchers and auditors should not only be granted access to all publicly accessible content on the platform (in anonymized or pseudonymized form), including content removed for terms of service violations[22] (see Box 3).[23] Where justified for the purpose of a research project, they should also receive access to technical documentation related to

[19] Matamoros-Fernández and Farkas, "Racism, Hate Speech, and Social Media"; Kayser-Bril, "Under the Twitter Streetlight."
[20] Pappas, "Strengthening Our Commitment to Transparency"; Roth and Gadde, "Expanding Access beyond Information Operations"; "YouTube Research."
[21] For a more extensive discussion of some of these aspects, see, for example, Vermeulen, "The Keys to the Kingdom"; Ausloos, Leerssen, and ten Thije, "Operationalizing Research Access in Platform Governance - What to Learn from Other Industries?"; Leerssen, "The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems"; for a discussion of regulatory and third-party audits in particular, see Ada Lovelace Institute, "Technical Methods for Regulatory Inspection of Algorithmic Systems"; Ada Lovelace Institute, "Inspecting Algorithms in Social Media Platforms"; Wagner et al., "Auditing Big Tech: Combating Disinformation with Reliable Transparency."
[22] Note that this should not include illegal content.
[23] Additionally, where data access is administered via APIs, platforms should refrain from constraining researchers' ability to carry out their work by imposing unnecessarily restrictive rate limits.

platforms' recommender systems and be able to review internal policies and processes related to recommender systems (e.g., on content demotion or promotion). Additionally, researchers should be provided with access to recommender systems (for example via APIs) or other tools allowing them to conduct tests or simulate recommendations and user pathways.

At the same time, **data access should be conditional on adhering to data protection rules and principles as well as robust security protocols**. The European Digital Media Observatory's proposed code of conduct on researcher access to platform data extensively discusses relevant concerns and safeguards in this regard.[24] Additionally, data that warrants particularly strong safeguards (such as data posing a risk of re-identifying a user) could, for example, be accessed through so-called "clean rooms"; that is, secured and strictly monitored research environments from which no data may be taken without prior privacy review.[25] Such an approach has also been put forward in the proposed Platform Accountability and Transparency Act (PATA) in the U.S..[26]

Box 3

Data access for vetted researchers

Data about publicly accessible content made accessible to vetted researchers could include the information:

- Content metadata (e.g., format, language, description, geographic origin, time of publication)
- Information on whether content was removed, the time of removal, and the reason for removal
- Information on whether content was reported and/or considered for removal
- Information on whether content was demoted and the reason for demotion
- Information on whether content was promoted by the platform and under what policy
- Information on how often content was recommended, viewed, and interacted with

[24] European Digital Media Observatory and Institute for Data, Democracy & Politics, George Washington University, "Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access"; for a discussion of additional concerns and trade-offs in this respect, see Keller, "User Privacy vs. Platform Transparency."
[25] Persily, "A Proposal for Researcher Access to Platform Data."
[26] Coons, Portman, and Klobuchar, Platform Accountability and Transparency Act.

Beyond enabling data access for researchers and auditors, it is also important to allow for research that does not rely on platforms as a potential gatekeeper. However, researchers who use their own tools to study platforms, create so-called "sock puppet" accounts for research purposes, or crowdsource or scrape data often are at risk of facing legal action from platforms, resulting in chilling effects on such research.[27] For this reason, **a safe harbor should be created for researchers, civil society organizations, and journalists conducting public interest research** in compliance with research ethics and data protection rules and principles.[28] The approach put forward in the proposed Platform Accountability and Transparency Act in the U.S. would be a step in the right direction in this regard. Further, amending existing legislation, like the U.S. Computer Fraud and Abuse Act (CFAA), to include explicit protections for public interest research could provide researchers with more legal certainty.

Audit recommender systems

Our recommendations above would create the foundation for platforms to be held accountable by researchers, journalists, and the public at large. But platforms should not only be held accountable by others — they need to carry out their own due diligence. To this end, **platforms should conduct or commission audits of their main recommender systems to systematically identify and manage potential risks and manifested harms** to individuals and the public interest that arise from the design, functioning, and use of their recommender systems.[29] This can help platforms better price into their operations the externalities caused by their services and better take stock not only of business and reputational risk, but also of societal risks and risks to users of their service.

For such an exercise to be meaningful and not a box-ticking exercise, audits should ideally meet two key criteria: first, they should be conducted by sufficiently independent third-party auditors with the necessary access to systems, data, and documentation; second, audit reports should be published — without divulging sensitive information — to allow for scrutiny from experts and the general public. If necessary, such audits could also be mandated by regulators for the largest online platforms. A similar approach has been adopted by the DSA in the EU and early

---

[27] See, for example, recent controversies around Meta shutting down research on Instagram and Facebook: Brandom, "Facebook Shut down German Research on Instagram Algorithm, Researchers Say"; Edelson and McCoy, "We Research Misinformation on Facebook. It Just Disabled Our Accounts."; Erwin, "Why Facebook's Claims about the Ad Observer Are Wrong."

[28] For additional context, see Hansen Shapiro et al., "New Approaches to Platform Data Research"; Heldt, Kettermann, and Leerssen, "The Sorrows of Scraping for Science"; Abdo et al., "A Safe Harbor for Platform Research."

[29] For a discussion of algorithmic impact assessments and internal algorithmic audits, see Moss et al., "Assembling Accountability"; Raji et al., "Closing the AI Accountability Gap."

lessons from its application as well as new insights from the emerging field of algorithmic auditing will provide valuable lessons on what an effective framework for audits of platform recommender systems can look like.[30]

Box 4

Summary of recommendations in this section

- Disclose detailed information on how recommender systems work and how they are operated, made accessible on a designated information page.
- Provide detailed information about content demotion (or reduction) policies, including criteria for when content is demoted and descriptions of how such policies are implemented.
- Report aggregate data on demoted content as part of existing transparency reporting formats.
- Publish information on and provide tools to analyze the content, pages, creators, and links that garner the most views and engagement.
- Provide qualified independent researchers conducting research in the public interest with access to data and documentation while adhering to data protection rules and principles and robust security protocols.
- Create a safe harbor for researchers, civil society organizations, and journalists conducting public interest research in compliance with research ethics and data protection rules and principles.
- Audit recommender systems and publish audit reports.

B. Empowered and informed users

Too often, platforms treat users as passive consumers rather than active participants in their online experience. This paternalistic attitude towards users manifests in different ways: sometimes, users lack the means to exercise meaningful control over their experience on the platform; other times, they lack basic information from platforms needed to make informed choices. But users should not be denied agency. The following recommendations therefore aim at empowering and informing users and enabling them to better shape their experience online.

[30] For a discussion of the risk assessments proposed in the EU's draft Digital Services Act, see Vermeulen, "The Keys to the Kingdom."

Give users more control over recommendations

First, **platforms should allow users to better customize the recommendations or content displayed to them**.[31] For instance, providing additional controls or choices can enable users to better protect their safety on the platform. To this end, users should generally be able to exclude certain keywords, creators, or pages from their recommendations. Further, they should be able to indicate certain types of content they want to exclude from recommendations or be less exposed to (for example through a "do not recommend" button). While many platforms have already implemented such controls, recent Mozilla research on YouTube user controls has also shown that merely offering such controls is not equivalent to effectively empowering users.[32] Instead, the research has found that YouTube's user controls are both confusing and often fail to produce the purported effect, only insufficiently or temporarily suppressing unwanted recommendations.[33] This example illustrates that controls (and intended outcomes) should be easy to understand and access as well as effective at what they purport to achieve.

Platforms should further **enable users to exert better control over which of their data, including personal data, is collected and subsequently used to inform recommendations**.[34] This should also extend to being able to exclude data collected from other related products or services (e.g., Google search data used to inform YouTube recommendations) or from previous engagement with certain content, pages, or users. In all cases, such controls should be designed in a manner that promotes genuine choice rather than using deceptive design practices incentivizing users to choose the platforms' preferred outcome.

Further, **users should be provided with an option to opt out of personalized recommendations or to influence the degree of personalization at any point in their product experience**. Personalization in this context means recommendations based on personal or behavioral information related to users that was observed or inferred by platforms (i.e., platform-driven personalization). It should not extend to user-driven personalization, such as following or subscribing to certain accounts or channels. Should users opt out of personalization, they might, for example, receive chronological

---

[31] See also Singh, "Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds"; Singh, "Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences"; additionally, Harambam et al. provide an interesting perspective on user attitudes towards enhanced control over recommendations in "Designing for the Better by Taking Users into Account."

[32] Ricks and McCrosky, "Does This Button Work?"

[33] In fact, an alternative user interface tested by Mozilla proved far more effective than the controls provided by YouTube.

[34] See also, for example, Ranking Digital Rights, "2020 Ranking Digital Rights Corporate Accountability Index: Recommendations."

recommendations, recommendations including the most popular content in a given location, or purely search term-based recommendations instead.

Enable users to make more informed choices

While most platforms do provide some transparency to users into why specific advertisements are presented to them, they do little in this regard when it comes to regular content. But having some understanding of why certain content is displayed can be important for users to better customize their experience on the platform — also given that empirical evidence suggests that many users still lack awareness of the algorithmic curation of their feeds.[35] Despite limitations of AI explainability, **platforms should do their best to provide users with an easily accessible and meaningful explanation of why specific content is displayed to them**. This should at a minimum include information on the key data points or signals influencing the recommendation in question. Furthermore, the explanation should point to information on how the recommender system works more generally (see above) and to controls available to users.

Box 5

Summary of recommendations in this section

- Allow users to better customize the recommendations or content displayed to them, by providing them with more and more effective controls.
- Enable users to exert better control over which of their data, including personal data, is collected and subsequently used to inform recommendations.
- Allow users to opt out of personalized recommendations or to influence the degree of personalization at any point in their product experience.
- Provide users with easily accessible and meaningful explanations of why specific content is displayed to them.

Finally, like content and account removals, content demotion can have significant impacts on people's ability to participate in the public conversation and express opinions. Moreover, it can also cause adverse economic impacts, for example for professional creators or artists who rely on platforms for distribution or monetization. But while users are generally notified by platforms when their content or account is

[35] Eslami et al., "'I Always Assumed That I Wasn't Really That Close to [Her]': Reasoning about Invisible Algorithms in News Feeds"; Gran, Booth, and Bucher, "To Be or Not to Be Algorithm Aware"; Kozyreva et al., "Public Attitudes towards Algorithmic Personalization and Use of Personal Data Online."

removed, content is demoted silently. As a general rule, and in the spirit of the Santa Clara Principles on Transparency and Accountability in Content Moderation,[36] platforms should also notify users when their content is demoted as a content moderation intervention.[37] Notices should include information identifying the content in questions as well as the reasons for which it was demoted (e.g., because the content was considered to be close to violating a platform's nudity policy). After all, in most cases, this relates to content that is expressly permitted on the platform.

Conclusion

Investigating the role large online platforms' recommender systems play in shaping individual and collective experiences — both online and offline — comes with great challenges. In the absence of better insights into how these systems function and better access to data from or about platforms, their impact remains poorly understood. At the same time, many harms caused or added to by platform recommender systems are already well established. It is therefore critical to both rein in these harms while creating the foundation for a better understanding and higher scrutiny of these systems and the companies that deploy them.

This paper seeks to provide a snapshot in time in this fast-moving field and debate. It does not and cannot offer an exhaustive list and discussion of potential options aiming to tackle the challenges we face relating to platform recommender systems. It does, however, lay out the steps Mozilla thinks are suitable and necessary at this point in time to build a better recommendation ecosystem. Our proposal is systemic in nature, prioritizing a concert of actions over quick individual fixes. Our goal is to hold platforms accountable while providing the public, experts, and individuals with bespoke tools to scrutinize recommendations and the engines driving them.

At the same time, more transparency, accountability, and control don't obviate the need for thinking more deeply about how recommender systems can be designed in a way that is more aligned with the public interest. Exciting work on this question and others is emerging and close attention should be paid to it going forward as our understanding of the issue space will evolve.[38]

---

[36] "Santa Clara Principles on Transparency and Accountability in Content Moderation."

[37] Some exceptions might be justified, for example in the case of accounts demonstrably used for spamming.

[38] See, for example, Ovadya, "Bridging-Based Ranking"; Goodman, "Digital Information Fidelity and Friction"; Massachi, "How to Save Our Social Media by Treating It like a City"; Stray, "Aligning AI Optimization to Community Well-Being"; Stray, "Beyond Engagement: Aligning Algorithmic Recommendations With Prosocial Goals."

Ultimately, our contribution to this discussion should be seen as an invitation to engage and provide feedback in order to advance the debate and move towards a consensus of what is necessary.

References

Abdo, Alex, Ramya Krishnan, Stephanie Krent, Evan Welber Falcón, and Andrew Keane Woods. "A Safe Harbor for Platform Research." Knight First Amendment Institute, 2022. http://knightcolumbia.org/content/a-safe-harbor-for-platform-research.

Ada Lovelace Instiute. "Inspecting Algorithms in Social Media Platforms," 2020. https://www.adalovelaceinstitute.org/report/inspecting-algorithms-in-social-media-platforms/.

———. "Technical Methods for Regulatory Inspection of Algorithmic Systems," 2021. https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/.

Alba, Davey. "Meta Pulls Support for Tool Used to Keep Misinformation in Check." *Bloomberg*, June 23, 2022. https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check.

Ausloos, Jeff, Paddy Leerssen, and Pim ten Thije. "Operationalizing Research Access in Platform Governance - What to Learn from Other Industries?" Algorithm Watch, 2020. https://algorithmwatch.org/de/wp-content/uploads/2020/06/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf.

Bösch, Marcus, and Becca Ricks. "Broken Promises: TikTok and the German Election." Mozilla, 2021. https://foundation.mozilla.org/en/campaigns/tiktok-german-election-2021/.

Brandom, Russell. "Facebook Shut down German Research on Instagram Algorithm, Researchers Say." The Verge, August 13, 2021. https://www.theverge.com/2021/8/13/22623354/facebook-instagram-algorithm-watch-research-legal-threat.

Brandt, Jessica, Bret Schafer, Elen Aghekyan, Valerie Wirtschafter, and Adya Danaditya. "Winning the Web: How Beijing Exploits Search Results to Shape Views of Xinjiang and COVID-19." The Brookings Institution, May 27, 2022. https://www.brookings.edu/wp-content/uploads/2022/05/FP_20220525_china_seo_v2.pdf.

Brown, Deborah, and Maria Laura Canineu. "Social Media Platforms Are Failing Brazil's Voters." *Human Rights Watch* (blog), October 28, 2022. https://www.hrw.org/news/2022/10/28/social-media-platforms-are-failing-brazils-voters.

Coons, Chris, Rob Portman, and Amy Klobuchar. Platform Accountability and Transparency Act (2021).

https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf.

Edelson, Laura, and Damon McCoy. "We Research Misinformation on Facebook. It Just Disabled Our Accounts." *The New York Times*, August 10, 2021, sec. Opinion. https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html.

Erwin, Marshall. "Why Facebook's Claims about the Ad Observer Are Wrong." *Mozilla* (blog). Accessed August 4, 2022. https://blog.mozilla.org/en/mozilla/news/why-facebooks-claims-about-the-ad-observer-are-wrong/.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "'I Always Assumed That I Wasn't Really That Close to [Her]': Reasoning about Invisible Algorithms in News Feeds." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–62. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015. https://doi.org/10.1145/2702123.2702556.

European Digital Media Observatory and Institute for Data, Democracy & Politics, George Washington University. "Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access," 2022. https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf.

Eusebio, Gerardo. "Fake News and Internet Propaganda, and the Philippine Elections: 2022." *Rappler*, May 23, 2022, sec. Philippine elections. https://www.rappler.com/technology/features/analysis-fake-news-internet-propaganda-2022-philippine-elections/.

Gillespie, Tarleton. "Reduction / Borderline Content / Shadowbanning." In *Platform Governance Terminologies Essay Series*. Yale-Wikimedia Initiative on Intermediaries & Information, 2022. https://law.yale.edu/sites/default/files/area/center/isp/documents/reduction_ispessayseries_jul2022.pdf.

Goodman, Ellen P. "Digital Information Fidelity and Friction." *Knight First Amendment Institute* (blog), 2020. https://knightcolumbia.org/content/digital-fidelity-and-friction.

Goodrow, Christos. "On YouTube's Recommendation System." *YouTube Official Blog* (blog), 2021. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

Gran, Anne-Britt, Peter Booth, and Taina Bucher. "To Be or Not to Be Algorithm Aware: A Question of a New Digital Divide?" *Information, Communication & Society* 24, no. 12 (September 10, 2021): 1779–96.

https://doi.org/10.1080/1369118X.2020.1736124.

Haimson, Oliver L., Daniel Delmonaco, Peipei Nie, and Andrea Wegner. "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 466:1-466:35. https://doi.org/10.1145/3479610.

Hansen Shapiro, Elizabeth, Michael Sugarman, Fernando Bermejo, and Ethan Zuckerman. "New Approaches to Platform Data Research." Netgain Partnership, 2021. https://drive.google.com/file/d/1bPsMbaBXAROUYVesaN3dCtfaZpXZgI0x/view?usp=sharing&usp=embed_facebook.

Harambam, Jaron, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. "Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems." In *Proceedings of the 13th ACM Conference on Recommender Systems*, 69–77. RecSys '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3298689.3347014.

Heldt, Amélie. "Borderline Speech: Caught in a Free Speech Limbo?" Internet Policy Review, October 15, 2020. https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510.

Heldt, Amélie, Matthias C. Kettermann, and Paddy Leerssen. "The Sorrows of Scraping for Science." *Verfassungsblog* (blog), 2020. https://verfassungsblog.de/the-sorrows-of-scraping-for-science/.

Integrity Institute. "Widely Viewed Content Dashboard." Integrity Institute. Accessed August 3, 2022. http://datastudio.google.com/reporting/28bc32fd-a067-4b4a-9be0-637e8c9bd917.

Jeantet, Diane. "Brazilian Voters Bombarded with Misinformation before Vote." *Associated Press*, October 25, 2022, sec. Jair Bolsonaro. https://apnews.com/article/jair-bolsonaro-caribbean-social-media-brazil-b75e51f515b0a708800198afa1b8601d.

Kayser-Bril, Nicolas. "Under the Twitter Streetlight: How Data Scarcity Distorts Research." *AlgorithmWatch* (blog), 2020. https://algorithmwatch.org/en/data-access-researchers-left-on-read/.

Keller, Daphne. "Amplification and Its Discontents." Knight First Amendment Institute, 2020. https://knightcolumbia.org/content/amplification-and-its-discontents.

———. "User Privacy vs. Platform Transparency: The Conflicts Are Real and We Need

to Talk About Them." *Stanford Law School Center for Internet and Society* (blog), 2022. https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0.

Klepper, David. "As 2022 Midterms Approach, Disinformation on Social Media Platforms Continues." PBS NewsHour, October 21, 2022. https://www.pbs.org/newshour/politics/as-2022-midterms-approach-disinformation-on-social-media-platforms-continues.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Ralph Hertwig, Stephan Lewandowsky, and Stefan M. Herzog. "Public Attitudes towards Algorithmic Personalization and Use of Personal Data Online: Evidence from Germany, Great Britain, and the United States." *Humanities and Social Sciences Communications* 8, no. 1 (2021): 1–11. https://doi.org/10.1057/s41599-021-00787-w.

Leerssen, Paddy. "The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems." *European Journal of Law and Technology* 11, no. 2 (2020). https://doi.org/10.2139/ssrn.3544009.

Little, Olivia, and Abbie Richards. "TikTok's Algorithm Leads Users from Transphobic Videos to Far-Right Rabbit Holes." Media Matters for America, 2021. https://www.mediamatters.org/tiktok/tiktoks-algorithm-leads-users-transphobic-videos-far-right-rabbit-holes.

Madung, Odanga. "Exporting Disinformation: How Foreign Groups Peddle Influence in Kenya through Twitter." Mozilla, 2022. https://foundation.mozilla.org/en/campaigns/exporting-disinformation-how-foreign-groups-peddle-influence-in-kenya-through-twitter/.

———. "From Dance App to Political Mercenary: How Disinformation on TikTok Gaslights Political Tensions in Kenya." Mozilla, 2022. https://foundation.mozilla.org/en/campaigns/kenya-tiktok/.

———. "Opaque and Overstretched, Part II: How Platforms Failed to Curb Misinformation during the Kenyan 2022 Election." Mozilla Foundation, 2022. https://foundation.mozilla.org/en/campaigns/opaque-and-overstretched-part-ii/.

Martiny, Clara, Isabel Jones, and Lucy Cooper. "Election Disinformation Thrives Following Social Media Platforms' Shift to Short-Form Video Content." *Institute for Strategic Dialogue* (blog), 2022. https://www.isdglobal.org/digital_dispatches/election-disinformation-thrives-following-social-media-platforms-shift-to-short-form-video-content/.

Massachi, Sahar. "How to Save Our Social Media by Treating It like a City." MIT Technology Review, 2021. https://www.technologyreview.com/2021/12/20/1042709/how-to-save-social-media-treat-it-like-a-city/.

Matamoros-Fernández, Ariadna, and Johan Farkas. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique." *Television & New Media* 22, no. 2 (February 1, 2021): 205–24. https://doi.org/10.1177/1527476420982230.

McCrosky, Jesse, and Brandi Geurkink. "YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm." Mozilla, 2021. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

Meta. "Widely Viewed Content Report: What People See on Facebook." Accessed August 10, 2022. https://transparency.fb.com/data/widely-viewed-content-report/.

Moss, Emmanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." Data & Society, 2021. https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf.

Mosseri, Adam. "Shedding More Light on How Instagram Works." *Instagram* (blog), 2021. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works.

Mozilla Foundation. "Facebook: Stop Group Recommendations." Accessed November 30, 2022. https://foundation.mozilla.org/en/campaigns/facebook-stop-group-recommendations/.

Mozilla Foundation. "Tell Twitter to Pause Trends until US Election Results Are Certified." Accessed November 30, 2022. https://foundation.mozilla.org/en/campaigns/untrendtwitter/.

Nicholas, Gabriel. "Shedding Light on Shadowbanning." Center for Democracy and Technology, 2022. https://cdt.org/insights/shedding-light-on-shadowbanning/.

Ovadya, Aviv. "Bridging-Based Ranking." Harvard Kennedy School Belfer Center for Science and International Affairs, 2022. https://www.belfercenter.org/publication/bridging-based-ranking.

Pappas, Vanessa. "Strengthening Our Commitment to Transparency." *Newsroom | TikTok* (blog), 2022. https://newsroom.tiktok.com/en-us/strengthening-our-commitment-to-transparency.

Persily, Nathaniel. "A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act." *Journal of Online Trust and Safety* 1, no. 1 (October 28, 2021). https://doi.org/10.54501/jots.v1i1.22.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit

Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." arXiv, January 3, 2020. https://doi.org/10.48550/arXiv.2001.00973.

Raji, Inioluwa Deborah, and Jingying Yang. "ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles." arXiv, January 7, 2020. https://doi.org/10.48550/arXiv.1912.06166.

Ranking Digital Rights. "2020 Ranking Digital Rights Corporate Accountability Index: Recommendations." Accessed January 10, 2022. https://rankingdigitalrights.org/index2020/recommendations.

Richards, Abbie. "Examining White Supremacist and Militant Accelerationism Trends on TikTok." *GNET* (blog). Accessed August 2, 2022. https://gnet-research.org/2022/07/18/examining-white-supremacist-and-militant-accelerationism-trends-on-tiktok/.

Ricks, Becca, and Jesse McCrosky. "Does This Button Work? Investigating YouTube's Ineffective User Controls." Mozilla Foundation, 2022. https://foundation.mozilla.org/en/research/library/user-controls/report/.

Roose, Kevin. "Inside Facebook's Data Wars." *The New York Times*, July 14, 2021, sec. Technology. https://www.nytimes.com/2021/07/14/technology/facebook-data.html.

Roth, Yoel, and Gadde. "Expanding Access beyond Information Operations," 2021. https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-.

Santa Clara Principles. "Santa Clara Principles on Transparency and Accountability in Content Moderation." Accessed August 10, 2022. https://santaclaraprinciples.org/images/santa-clara-OG.png.

Shenkman, Carey, Dhanaraj Thakur, and Emma Llansó. "Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis." Center for Democracy and Technology, 2021. https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/.

Singh, Spandana. "Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds." New America, 2019. http://newamerica.org/oti/reports/rising-through-ranks/.

———. "Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences." New America, 2020. http://newamerica.org/oti/reports/why-am-i-seeing-this/.

Stanley-Becker, Isaac, and Drew Harwell. "Misinformation Floods the Midterms, at Times Urging Violence." *Washington Post*, November 8, 2022.

https://www.washingtonpost.com/elections/2022/11/08/election-day-disinformation/.

Stepanov, Anna. "Sharing Our Content Distribution Guidelines." *Meta* (blog), September 23, 2021. https://about.fb.com/news/2021/09/content-distribution-guidelines/.

Stray, Jonathan. "Aligning AI Optimization to Community Well-Being." *International Journal of Community Well-Being* 3, no. 4 (December 1, 2020): 443–63. https://doi.org/10.1007/s42413-020-00086-3.

———. "Beyond Engagement: Aligning Algorithmic Recommendations With Prosocial Goals." *Partnership on AI* (blog), 2021. https://partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals/.

The Virality Project. "Memes, Magnets and Microchips: Narrative Dynamics around COVID-19 Vaccines." Stanford Digital Repository., 2022. https://purl.stanford.edu/mx395xj8490.

Thomas, Elise, and Kata Balint. "Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere.'" Reset, Institute for Strategic Dialogue, 2022. https://www.isdglobal.org/wp-content/uploads/2022/04/Algorithms-as-a-weapon-against-women-ISD-RESET.pdf.

Thorburn, Luke, Priyanjana Bengani, and Jonathan Stray. "How Platform Recommenders Work." *Understanding Recommenders* (blog), February 8, 2022. https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a.

TikTok. "How TikTok Recommends Videos #ForYou." *Newsroom | TikTok* (blog), 2020. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.

University of the Philippines. "Filipino Voters Were Engulfed in Relentless Stream of Disinformation," May 11, 2022. https://up.edu.ph/filipino-voters-were-engulfed-in-relentless-stream-of-disinformation/.

Vermeulen, Mathias. "The Keys to the Kingdom." Data and Democracy. Knight First Amendment Institute, 2021. https://knightcolumbia.org/content/the-keys-to-the-kingdom.

Wagner, Ben, Johanne Kübler, Lubos Kuklis, and Carolina Ferro. "Auditing Big Tech: Combating Disinformation with Reliable Transparency." Tallinn, Estonia: Enabling Digital Rights and Governance & Omidyar Network, 2021. https://enabling-digital.eu/wp-content/uploads/2021/02/Auditing_big_tech_Final.pdf.

Wall Street Journal. "The Facebook Files," October 1, 2021, sec. Tech.

https://www.wsj.com/articles/the-facebook-files-11631713039.
YouTube. “YouTube Research.” Accessed August 10, 2022. https://research.youtube/.