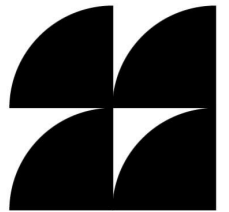


MARCH 2025

Better Feeds: Algorithms That Put People First

*A How-To Guide for Platforms and
Policymakers*

KGI EXPERT REPORT



PLTF EX

PLTF EX-07557

PLTF EX-07557_0001

KGI Expert Working Group on Recommender Systems

This report is the product of the [KGI Expert Working Group on Recommender Systems](#), whose members are listed below. KGI convenes expert working groups (EWGs) that bring together relevant experts for a time-bound project to summarize knowledge and articulate policy options. KGI builds EWGs by inviting selected volunteer contributors from across academia, industry, civil society, journalism, and practitioner communities to engage in a focused collaboration on a specific policy topic over the course of 6-12 months or more. The goal of each EWG is to produce non-partisan resources of broad utility to policymakers and industry decisionmakers. Learn more about [KGI Expert Working Groups](#).

Alex Moehring
Purdue University

Leif Sigerson
Integrity Institute

Alissa Cooper
Knight-Georgetown Institute

Luke Thorburn
King's College London

Arvind Narayanan
Princeton University

Matt Motyl
Psychology of Technology Institute, University of Southern California

Aviv Ovadya
AI & Democracy Foundation

Motahhare Eslami
Carnegie Mellon University

Elissa Redmiles
Georgetown University

Nadine Farid Johnson
Knight First Amendment Institute at Columbia University

Jeff Allen
Integrity Institute

Nathaniel Lubin
Berkman Klein Center, Harvard University

Jonathan Stray
University of California, Berkeley

Ravi Iyer
University of Southern California Neely Center

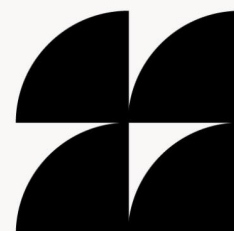
Julia Kamin
Prosocial Design Network

Zander Arnao
Knight-Georgetown Institute



About the Knight-Georgetown Institute

The Knight-Georgetown Institute (KGI) is dedicated to connecting independent research with technology policy and design. KGI serves as a central hub for the growing network of scholarship that seeks to shape how technology is used to produce, disseminate, and access information. KGI is designed to provide practical resources that policymakers, journalists, and private and public sector leaders can use to tackle information and technology issues in real time. Georgetown University and the Knight Foundation came together to launch the institute in 2024. Learn more about KGI at <https://kgi.georgetown.edu>.



Executive Summary

The algorithmic recommender systems that select, filter, and personalize experiences across online platforms and services play a significant role in shaping user experiences online. These systems largely determine what users see, read, and watch, fueling debates around their potential to amplify harmful content, foster societal division, and prioritize engagement over user well-being. In reaction, some policymakers have turned to blanket bans on personalization or to the promotion of chronological feeds. But there are many better alternatives. Suggesting that users must choose between today's default feeds and chronological or non-personalized feeds creates a false choice.

This report, prepared by the KGI Expert Working Group on Recommender Systems, offers comprehensive insights and policy guidance aimed at optimizing recommender systems for long-term user value and high-quality experiences. Drawing on a multidisciplinary research base and industry expertise, the report highlights key challenges in the current design and regulation of recommender systems and proposes actionable solutions for policymakers and product designers.

A key concern is that some platforms optimize their recommender systems to maximize certain forms of predicted engagement, which can prioritize clicks and likes over stronger signals of long-term user value. Maximizing the chances that users will click, like, share, and view content this week, this month, and this quarter aligns well with the business interests of tech platforms monetized through advertising. Product teams are rewarded for showing short-term gains in platform usage, and financial markets and investors reward companies that can deliver large audiences to advertisers.

Concerns have been raised about the relationship between this design approach and a range of individual and societal harms, including the spread of low-quality or harmful information, reduced user satisfaction, problematic overuse, and increased polarization. Available evidence underscores the need for a shift towards designs that optimize for long-term user satisfaction, well-being, and societal benefits.

To achieve this, the KGI Expert Working Group on Recommender Systems proposes that policymakers and product designers adopt the following:

- **Detailed transparency** in the design of recommender systems, including the public disclosure of input data sources, value model weights, and metrics used to measure long-term user value. Platforms must also publicly disclose the internal metrics used to assess product teams responsible for recommender system design.
- **User choices and defaults** that allow individuals to tailor their platform experiences and switch between different recommendation systems. Minors must be provided with default recommender systems optimized to deliver them long-term value.

- **Assessments of long-term impact**, where platforms continuously test the impact of algorithmic changes over extended periods. Platforms must conduct these assessments by running so-called “holdout” experiments that exempt a group of users from design changes for 12 months or more. Public disclosure of aggregated experiment results and independent audits must be adopted for accountability.

This report provides a how-to guide for the implementation of each set of proposals, lighting a pathway towards higher quality designs that may still be personalized or leverage some forms of engagement data, but overcome the design flaws of engagement-optimized systems.

By following this expert working group’s guidance, summarized below, platforms and policymakers can help to address the harms associated with recommender systems while preserving their potential to enhance user experiences and societal value. This report serves as a roadmap for any policymaker or product designer interested in promoting algorithmic systems that put users’ long-term interests front and center.

Core Policy Guidance¹	
Design Transparency	Platforms must publicly disclose information about the specific input data and weights used in the design of their recommender systems.
	Platforms must publicly disclose the metrics they use to measure long-term user value.
	Platforms must publicly disclose the metrics they use to evaluate product teams responsible for recommender system design.
User Choices and Defaults	Platforms must offer users an easily accessible choice of different recommender systems. At least one of these choices must be optimized to support long-term value to users.
	Platforms must provide easily accessible ways for users to set their preferences about types of items to be recommended and to be blocked. Platforms must honor those preferences.
	By default, platforms must set minors’ recommender systems to be optimized to support long-term value to these users. If platforms have insufficient information about long-term value to minors, they must default to non-personalized recommender systems.

¹ The Core Policy Guidance was designed with the US legal framework in mind.

Long-Term Holdout Experiments	Platforms must run long-term (12-month or longer) holdout experiments on a continuous basis.
	Platforms must report the aggregate, anonymized results of the holdout experiments publicly.
	Holdout experiments must be subject to an audit by an independent third party.
Global Policy Guidance²	
Public Content Transparency	Platforms must continuously publish a sample of the public content that is most highly disseminated on the platform and a sample of the public content that receives the highest engagement.
	Platforms must continuously publish a representative sample of public content consumed during a typical user session on the platform at any given time.
User Defaults	By default, platforms must optimize users' recommender systems to support long-term user value.
Metrics and Measurement	Platforms must measure the aggregate harms to at-risk populations that result from recommender systems and publicly disclose the results of those measurements.

² This is additional guidance tailored for implementation in jurisdictions outside the US.

Table of Contents

- I. Introduction.....1**
- II. Background: Recommender Systems 101.....4**
 - A. Terminology.....4
 - B. Designing Recommender Systems.....5
 - C. Understanding Signals and Predictions.....7
 - D. Characterizing User Signals.....10
 - E. Maximizing Predicted Engagement.....12
- III. Research Findings.....13**
 - A. Harms Associated with Maximizing Predicted Engagement.....13
 - B. Harms to Minors.....15
 - C. User Preferences and Satisfaction.....16
 - D. Alternatives to Maximizing Predicted Engagement.....18
 - 1. Chronological and Non-Personalized Feeds.....18
 - 2. Better Approaches.....19
 - 3. Implications for Recommender System Design.....21
- IV. Policy Landscape.....23**
- V. Core Policy Guidance.....28**
 - A. Design Transparency.....30
 - B. User Choices and Defaults.....34
 - C. Long-Term Holdout Experiments.....38
- VI. Global Policy Guidance.....41**
 - A. Public Content Transparency.....41
 - B. User Defaults.....43
 - C. Metrics and Measurement.....44
- VII. Best Practices for Product Teams.....45**
 - A. How to Construct Holdout Groups.....45
 - B. How to Disclose Recommender System Weights.....46
- VIII. Conclusion.....47**
- Bibliography.....48**

I. Introduction

Every day, billions of people scroll through social media feeds, search results, and streaming recommendations that shape what they see, read, and watch. Algorithmic systems determine what to show each user, wielding enormous influence over our online experiences and, increasingly, our lives offline. While these recommendation algorithms have fueled some of the world's most successful businesses, they have also sparked intense debate about their role in amplifying harmful content, unwanted experiences, and societal division.

As the push for regulation of these systems intensifies, policymakers and product designers need evidence-informed solutions that move beyond the binary choice of chronological versus algorithmic feeds. This report examines a range of options, explains recommender systems in depth, and provides guidance informed by research for how to prioritize long-term value and high-quality experiences for users.

Algorithmic curation has become ubiquitous across social media, search, streaming services, e-commerce, gaming, and more. A single platform may deploy many different recommender systems, using them to power social media feeds, ad displays, comment sections, account recommendations, notifications, video and audio autoplay selections, curated home pages or landing pages, and many other features.

Maximizing the chances that users will click, like, share, and view content this week, this month, and this quarter aligns well with the business interests of tech platforms monetized through advertising. Product teams are rewarded for showing short-term gains in platform usage, and financial markets and investors reward companies that can deliver large audiences to advertisers.

But policymakers, advocates, and the public are increasingly drawing connections between the design of recommender systems and a variety of harms, including harms to adult or youth well-being (e.g., self-harm and eating disorders), fraud and scams, and civic or societal harms (for example, extremism and polarization). Some of these concerns center on how algorithms leverage “engagement” with content – actions taken by users such as clicking a link, liking a post, accepting a friend request, or playing a video. There is evidence that recommender systems that are overly reliant on certain forms of engagement can cause harm to individuals, communities, and society.

For example, in 2018, Meta CEO Mark Zuckerberg explained in an official company note how, at least on Facebook, the closer content is to violating the platform's policies (i.e., “borderline content”), the greater likelihood it is to receive engagement:

“Our research suggests that no matter where we draw the lines for what [content] is allowed, as a piece of content gets close to that line, people will engage with it more on average -- even

when they tell us afterwards they don't like the content...Interestingly, our research has found that this natural pattern of borderline content getting more engagement applies not only to news but to almost every category of content.”³

That borderline content evidently receives more engagement on platforms such as Facebook is likely a result of design choices made by their owners. Despite what Zuckerberg claimed, there is nothing “natural” about engagement patterns on social media: recommender systems play a central role in shaping the kinds of engagement the most highly ranked content elicits, such as by incentivizing content creators to shape their content in ways intended to garner more engagement.⁴ As a result, borderline content being disproportionately represented at the top of users’ feeds is not an inevitable outcome but rather a choice of design.

Concerns about the effects of algorithms designed in this way have fueled legislative activity, litigation, and product design changes in many jurisdictions. In the US, more than 75 bills were introduced across 35 states between 2023 and 2024 addressing social media algorithms; more than a dozen have been signed into law; and many of those were subsequently challenged in court. Many (but not all) of these bills aim specifically to protect youth online.

Lawsuits brought by state Attorneys General and private plaintiffs on behalf of individuals, children, families, and school districts have alleged a range of algorithm-related harms and have contributed to an ever-morphing body of case law as judges grapple with questions related to the First Amendment and Section 230 of the Communications Decency Act.

In the European Union, the landmark Digital Services Act (DSA) entered into force for the largest online platforms in 2023 and included provisions requiring specific recommender system designs and disclosures.

Many attempts to regulate recommender systems thus far have focused on restricting the use of algorithms, restricting personalization, or both.⁵ The motivation for these approaches is simple: if algorithmic curation, or personalized algorithmic curation, is viewed as the source of harm, then prohibiting the use of these algorithms or requiring users to opt in to use them seems like a good solution. In some cases, legislation has explicitly or implicitly endorsed chronological feeds – where all content available to the user is displayed in reverse chronological order – as a better alternative.

³ Zuckerberg, “A Blueprint for Content Governance and Enforcement.” This pattern was also demonstrated in internal Facebook documents disclosed by Frances Haugen. See Haugen, “Providing Negative Feedback Should Be Easy.”

⁴ See, e.g., Hödl and Myrach, “Content Creators Between Platform Control and User Autonomy”; Glotfelter, “Algorithmic Circulation”; Radesky et al., “Algorithmic Content Recommendations on a Video-Sharing Platform Used by Children.”

⁵ For example, bills introduced in many states would require platforms to be more transparent about algorithmic ranking of content and to permit users to opt out, usually by ranking content chronologically. See, e.g., Oklahoma, “Oklahoma Social Media Transparency Act”; Minnesota, “SF 2716.”

Blanket regulations that aim to prevent or limit the use of algorithmic recommender systems fail to account for the vast space of potential algorithmic designs that could be beneficial to users. Such broad policies miss a critical opportunity to require, incentivize, or guide recommender systems toward optimization for long-term user value and high-quality experiences – even when those optimizations rely to some extent on engagement, personalization, or both. Better recommender systems are possible.

Approaches to algorithm regulation thus far demonstrate a number of gaps and limitations:

- Some approaches treat all forms of engagement the same. A nuanced understanding of engagement shows that some forms of engagement provide higher value signals than others. For example, writing a long comment on another user’s post, filling out an in-feed survey, or making a purchase might all be considered engagement signals, yet they provide stronger, less ambiguous indications to the platform of a user’s preferences and intent than automatic or passive scrolling, liking, or clicking.
- Many approaches do not address the organizational incentives that drive design choices within companies, including the structures linking corporate, team, and employee objectives and metrics to rewards. Regulation can play a helpful role in changing these incentives to inspire designs that bring more value to users.
- Some approaches assume that the same strategy for mitigating harm will work on every platform and algorithmic feed. But platforms designed for different user engagement patterns and with different affordances require tailored approaches. For example, a platform that does not provide for “following” other accounts would not be able to meet a requirement to curate content solely from the user’s network of followed accounts.
- As noted above, some approaches hold up chronological feeds as the preferred alternative to algorithmic feeds. But chronological feeds have important limitations and naturally reward spam-like behavior. There are better options available that can be designed to mitigate a variety of harms.

This report offers a guide for policymakers and product designers to address these gaps and incentivize better recommender systems. It represents the consensus view of a leading group of recommender systems experts about measures that can be enacted through public policy and product design to optimize these systems for long-term user value and high-quality user experiences. Not every member of the KGI Expert Working Group on Recommender Systems agrees with every facet of this report, but as a collective the working group supports the guidance provided.

The guidance in this report was developed on the basis of a growing body of research spanning computer science, economics, social science, and behavioral science, combined with deep insights

from industry veterans. Following the guide would help mitigate many of the concerns currently motivating proposals for the regulation of online algorithms.

II. Background: Recommender Systems 101

Recommender systems are the algorithms that select, filter, and personalize content and other items across online platforms, services, and applications.⁶ This report considers recommender systems broadly, including those used across social media, search engines, streaming, e-commerce, and gaming. Although policy discussions about recommender systems tend to focus on social media, the policy guidance in this report may be equally applicable to recommender systems in use more broadly.

This section is intended to provide a primer on recommender systems for lay readers. It establishes terminology used throughout the rest of the report, describes how recommender systems are designed, explains the roles of engagement and personalization, and reviews different types of harms that have been associated with algorithmic systems.

A. Terminology

Candidate items are items identified during the ranking stage as plausibly of interest to a particular user in a particular context.⁷

Engagement refers to actions taken by users on recommended items, such as clicks, likes, comments, reposts, watch time, dwell time, upvote, downvote, and many others.⁸

Holdout groups are groups of users on a platform who are exempted from the application of changes to their user experiences for a fixed period of time.

Items are the elements eligible for display by a recommender system. Items can include individual pieces of content, accounts, groups, pages, channels, products, or ads. This report uses the term ‘item’ instead of ‘content’ because not everything that gets recommended to users is considered content in all circumstances (e.g., user accounts).

Long-term user value refers to outcomes that align with individual users’ deliberative, forward-looking preferences or aspirations. Long-term user value prioritizes long-run user preferences and users’ ability to achieve their aspirations over short-run, impulsive preferences.

Metrics are what is measured to evaluate the success of a recommender system at a high level.

⁶ Stray et al., “Building Human Values into Recommender Systems,” 1.

⁷ Thorburn et al., “How Platform Recommenders Work.”

⁸ Cunningham et al., “What We Know About Using Non-Engagement Signals in Content Ranking.”

Predictions are probabilities or other values generated by machine learning models that are weighted and combined in the value model used for ranking. These probabilities can be generated for any potential user behavior – click, like, watch, repost, upvote/downvote, etc.

Ranking is the stage of the recommendation process in which each candidate item is assigned a score intended to capture the value of showing it to a particular user in a particular context.⁹

Recommender systems are the algorithms that select, filter, and personalize content and other items across online platforms, services, and applications.

Scores are the output of value models: numeric values assigned to each candidate item during ranking. Computing a score for each candidate and then ordering them by their scores is the core function of ranking.

Signals are the data inputs used for ranking. They summarize aspects of the content, the user, the context, and how all of these interact.

Value models are the formulas used to compute a ranking score for an item for a particular user. Value models often have many terms, which are the individual components of the formula.

Weights are individual numeric settings that control the output of a recommender system at a high level, such as the relative contributions of different predictions to an item's score.¹⁰

B. Designing Recommender Systems

The process of recommending content proceeds from mapping the universe of items a user could plausibly be interested in to identifying and finally recommending the items most advantageous to a platform's goals. This process usually occurs in four stages:¹¹

1. **Moderation:** The platform applies its moderation policies to the universe of available items, removing items from the pool that violate the platform's policies.
2. **Candidate generation:** The platform selects high-potential items as candidates from among the universe of available items. This universe can be very large, sometimes on the order of billions of items, so selecting candidate items is usually a very lightweight computational process that does not involve in-depth analysis.

⁹ Ibid.

¹⁰ Stray et al., "Building Human Values into Recommender Systems," 13.

¹¹ Thorburn et al., "How Platform Recommenders Work."

3. **Ranking:** Each candidate item is then assigned a numeric score intended to capture the value of showing it to a particular user in a particular context.¹² This score determines the order in which candidate items are output from the ranking stage.
4. **Re-ranking:** Finally, the order of candidate items is changed according to other ancillary goals, for example, to avoid repetitiveness in terms of the content type or source. These ancillary goals may be considered important because their absence undermines long-term engagement – for example, less variety may be more engaging in the short term, but may diminish the user's experience on a platform over the long-term.

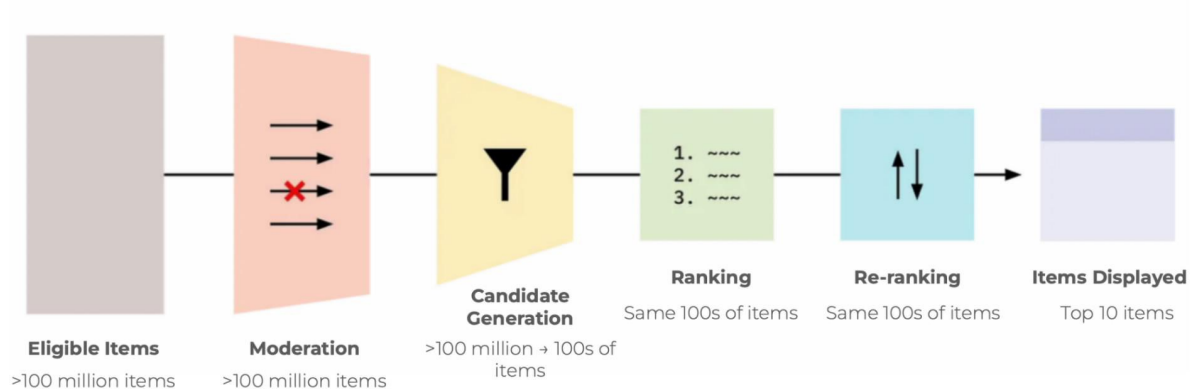


Figure 1. A typical recommender system pipeline, along with an approximate number of items retained at each stage for a large platform.¹³

There are multiple possible high-level designs for recommender systems. A traditional approach assigns fixed weights to specific predictions (such as probabilities of clicks, likes, or shares) based on their presumed importance to the user, and the system sums these terms to compute a score for each item.

A more complex recommender system may rely on a neural network (a type of machine learning model that mimics the human brain) to generate ranking scores instead. First, it relies on predefined rules to label past user behaviors, assigning a value to each behavioral signal (post, like, follow, etc.) based on the platform's criteria. This labeled data is then used to train a neural network, which learns to identify patterns and relationships between that user's specific behavior, the content characteristics, and the context. The neural network generalizes these patterns to predict the value of showing a specific item to a particular user at a given time, and the recommender system uses these predictions going forward.

Within a single online service or product, there may be separate recommender systems used for different feeds, such as the home page or main feed, sidebars, account or group recommendations,

¹² Ibid.

¹³ The graphic is modified with permissions from Thorburn et al., "How Platform Recommenders Work."

and ads. These recommender systems may be designed to accomplish different objectives.¹⁴ See Figure 2 below for examples based on recommendations from YouTube and Instagram.

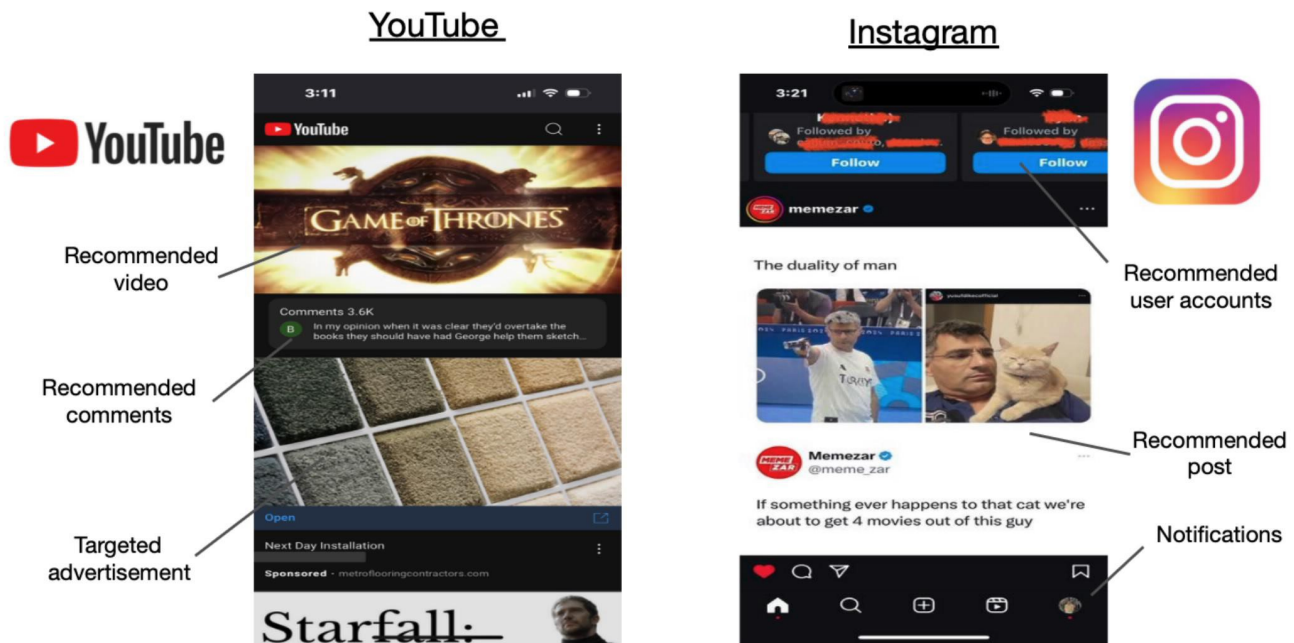


Figure 2. Social media platforms are designed as collections of many recommender systems. Above are two examples from YouTube and Instagram featuring recommender systems that rank different kinds of items.

C. Understanding Signals and Predictions

The type of signals and predictions used during the ranking stage are key choices in the overall design of recommender systems. As discussed below, most recommender systems make predictions about the likelihood of specific user behaviors resulting from a recommendation.¹⁵ However, predictions can also focus on the characteristics of individual items and the kinds of feedback users might give. The table below describes these categories of predictions in more detail. Platforms often use and combine predictions from all three categories.

¹⁴ See, e.g., Hosseinmardi et al., “Causally Estimating the Effect of YouTube’s Recommender System Using Counterfactual Bots.”

¹⁵ See Goodrow, “On YouTube’s recommendation system” as an example that provides an overview of the signals used by YouTube’s recommender system.

Categories of predictions**Engagement predictions**

The likelihood that a user will take an action on an item, such as a click, comment, like, re-share, watch, dwell or linger, or upvote/downvote.

Item-level scores¹⁶

A score assigned to an item independent of the viewer with a clear valence of value, that is, where a high score is more desirable and a low score is less desirable in some way. Item-level scores could include an item's informativeness, toxicity, or likelihood to be spam, for example. Item-level scores can be assigned by human raters, machine learning models trained on human-labeled data, or pre-defined heuristics (e.g., posts coming from a certain URL are automatically given a certain spam score). Some third parties have created indices containing scores for certain types of items, such as news credibility or authoritativeness.¹⁷

Item-level survey response predictions¹⁸

The likelihood of users to answer a particular item-level survey in a positive or negative way. See the discussion of signals below for more detail on how surveys can be incorporated into predictions.

The scope of possible signals recommender systems can incorporate is even wider. These signals can range from observations about behavior to qualitative data and explicit user controls. The table below describes various categories of potential signals by their source.

Categories of signals**Engagement (or behavioral) signals**

Actions a user can take on an item, such as clicks, comments, reactions, re-shares, watch time, dwell time, purchases, bookmarks, saves, and feature usage.

User responses to survey questions¹⁹

Users can take two types of surveys: those about specific items (item-level) and others about their overall experience, not associated with a specific item (user-level).

Examples of item-level questions: “Is this item informative?” “Would you like to see more like this?” “Is this item worth your time?” “What star rating would you give this item?”

Examples of user-level questions: “How was your experience using [platform] today?” “Would you use [platform] again?” “Would you recommend [platform] to a friend?”

¹⁶ Cunningham et al., “What We Know About Using Non-Engagement Signals,” 5.

¹⁷ See, e.g., NewsGuard, “News Reliability Ratings.”

¹⁸ Stray et al., “Building Human Values into Recommender Systems,” 18.

¹⁹ Cunningham et al., “What We Know About Using Non-Engagement Signals,” 2.

Quality feedback from users²⁰

User feedback on an item that has a quality valence (i.e., positive or negative). For example, platforms often include buttons allowing users to explicitly report items for violating platform rules or to request that they not be shown again.

Annotations from raters²¹

Annotations provided by human raters to identify particular kinds of ‘positive’ or ‘negative’ content for the purpose of ranking. For example, human raters could annotate a set of items for whether they reflect a constructive approach to disagreement or a diversity of viewpoints.

Content properties

Information about an item’s content, such as its media type (text, image, video, etc.) or higher order aggregations like topic and sentiment.

User controls²²

Settings that the user can explicitly and proactively change to control the ranking or visibility of future items. For example, functionalities that allow users to follow and subscribe to accounts, block users, mute terms, explicitly select topics, or change how posts are ranked. User controls are available in the user interface and their selection is not visible to other users.

Profile data

Data about the characteristics of users, such as their age, gender, race, ethnicity, religion, political affiliation, socioeconomic status, and location.

Off-platform data²³

Data collected about users’ activity from outside a platform’s recognizable suite of services, such as the list of URLs the user visits where the platform has placed tracking technologies (pixels, cookies, etc.).

Context data

Data that communicates contextual information about other data, such as date, time of day, and location.

²⁰ Ibid.

²¹ Stray et al., “Building Human Values into Recommender Systems,” 17.

²² Cunningham et al., “What We Know About Using Non-Engagement Signals,” 2.

²³ Various platforms collect data about users’ activity outside of their core suite of services. See, e.g., Meta, “Meta Pixel”; Reddit, “About the Reddit Pixel.” Many platforms have disclosed that this off-platform data is used by their on-platform recommender systems. See, e.g., Google, “Activity Controls”; Google, “Manage Your Recommendations and Search Results”; Meta, “Review Your Activity off Meta Technologies”; Reddit, “Settings.”

D. Characterizing User Signals

Users generate a wide variety of behavioral signals as they use mobile apps and visit websites, and not all signals are created equal. The amount, type, and value of information conveyed by any individual signal varies greatly depending on whether that signal is generated by liking a post, purchasing an item, toggling a user setting, or any other user activity.

Signals can be characterized on at least four different spectrums:

- **Impulsive to Deliberative:** This spectrum represents the user's state of mind when interacting with the recommender system, i.e., whether they are acting automatically (impulsively) or with a degree of intention or conscious self-reflection (deliberatively). Users could be acting impulsively when scrolling through their feeds for long periods of time, for example, and deliberately when writing long comments or selecting from among different videos.²⁴ Similar examples exist in other domains, for example, impulsively eating potato chips versus deliberately choosing a healthy meal.
- **Effortless to Onerous:** This spectrum represents the level of effort users expend when using different affordances or features of an online platform. Some actions on a platform require more time, focus, and steps to complete than others.
- **Inferred to Stated Preferences:** This spectrum represents the manner in which users express their preferences to platforms' recommender systems. In some cases, user preferences are inferred from their activity, while in others this process can be more explicit. For example, a platform might infer users' preferences based on the users dwelling on an item, or users may explicitly state their preferences when toggling user controls or responding to surveys.
- **Ambiguous to Clear:** This spectrum represents the confidence with which a given signal accurately captures information about what it purports to measure. For some signals there can be multiple interpretations about their meaning and therefore uncertainty as to whether they communicate meaningful information. For example, sometimes users comment on items they like, while at other times they comment on items they do not like, creating ambiguity about what the act of commenting (setting aside the content of the comment) indicates to the platform.

These spectrums are not entirely independent of one another. Some characteristics of signals will tend to correlate. For example, if a signal indicates a user behavior that occupies a lot of time (e.g., filling out a long survey), that onerous signal is also more likely to be deliberative.

²⁴ There can be individual variability in the online behaviors that signal impulsive versus deliberative usage. A user may post long comments when they are impulsively ranting, or users in a disorganized mental state may take a long time to look through different videos as they struggle to make a decision. Deliberative signals can thus be ambiguous.

The existence of specific signals can also be intertwined with the incentive structures of the platform. Content creators might encourage their followers to like a post or send a direct message because they perceive this behavior to be algorithmically beneficial to their content. Retailers might encourage users to leave reviews or reshare content to be entered into contests or sweepstakes.

Preventing the use of all user signals – as some policy proposals seek to do – ignores these nuanced distinctions between different types of signals. Heavier reliance on signals that are deliberative, onerous, clear, and reflect stated preferences should produce user experiences that are more aligned to users’ aspirations, in other words, feeds that are both personalized and valued by users. Similarly, if research validates that certain impulsive or effortless signals are better predictors than more deliberate, onerous signals, they could be used as well. The research behind these distinctions is explored further in Section III.

Examples of Each Signal Type

Spectrum	Examples
Impulsive $\longleftrightarrow$ Deliberative	Impulsive: resharing a link without clicking on it first, liking a video without watching most of it, writing short/quick comments Deliberative: writing long comments, filling out surveys, long watch time, bookmarks, user controls
Effortless $\longleftrightarrow$ Onerous	Effortless: clicks, reactions, reshares, upvotes/downvotes Onerous: writing long comments, filling out long surveys, long watch time
Inferred preference $\longleftrightarrow$ Stated preference	Inferred preference: dwell time, clicks, reshares Stated preference: user controls, survey responses, purchases, value-based reactions (e.g., “insightful”, “respect” buttons),²⁵ up-votes, searches for specific keywords

²⁵ LinkedIn, “Use LinkedIn Reactions”; Stroud et al., “Like, Recommend, or Respect?”; Washington Post, “The Washington Post Launches a New Commenting Experience Exclusively for Subscribers.”

Ambiguous $\longleftrightarrow$ Clear	Ambiguous: dwell time, watch time, clicks, likes, reshares, comments, feature usage, upvotes/downvotes, purchases (due to impulse buys), follows (due to tracking disfavored accounts) Clear: user controls, survey responses, value-based reactions, long-term user retention
---------------------------------------	--

E. Maximizing Predicted Engagement

Maximizing predicted engagement is among the most prevalent industry approaches for how recommender systems are designed to select and rank items. Engagement influences the signals, predictions, and metrics used by recommender systems.

Recommender systems today largely rely on signals, predictions, and metrics of engagement to determine which items to recommend. For example, a recommender system may use signals about which videos a user typically watches for long periods of time to generate predictions about which new videos will cause them to repeat this behavior.

Ad-supported platforms in particular have strong incentives to optimize their recommender systems for predicted engagement – more users spending more time and being more engaged means more ad revenue. Keeping users on a platform by maximizing their engagement also generates more data that allows a platform to improve its ad targeting, which incentivizes continued engagement maximization in a self-reinforcing cycle.²⁶

Many platforms have disclosed that their recommender systems are optimized for maximizing predicted engagement, for example:

- Facebook and Instagram’s recommender systems make predictions about how likely users are to tap, watch, and otherwise respond to recommended content, although predictions differ across media types (e.g., stories and posts).²⁷
- TikTok’s recommender systems have two key metrics – maximizing retention and time spent – and predict how likely a user will like, comment on, and watch posts.²⁸
- X’s recommender systems optimize for behaviors such as likes, reposts, and replies and assign ranking scores based on predictions about this engagement.²⁹ They also source roughly half of candidate items from accounts users follow and are likely to engage with, and half from items

²⁶ Wu, *The Attention Merchants*.

²⁷ Meta, “Our Approach to Explaining Ranking.”

²⁸ Smith, “How TikTok Reads Your Mind.”

²⁹ X, “Twitter’s Recommendation Algorithm.”

engaged with by users who have similar interests.

The design of recommender systems may be poorly aligned with prosocial values, such as safety and user empowerment.³⁰ This misalignment can be particularly concerning for minors. Public policy can work to create better incentives that are more consistent with delivering long-term value and high-quality experiences to all.

III. Research Findings

Recommender systems and the dominance of engagement-based designs have recently been the subject of significant research. This scholarship spans computer science, psychology, social science, and behavioral economics, and has gleaned many important insights for public policy. This section summarizes the major findings of this research and its lessons for policy affecting the design of recommender systems.

While the breadth of scholarship on recommender systems is wide, its insights for policy can be distilled into three major categories: documenting specific harms associated with engagement-maximizing design, examining the misalignment between recommendation strategies and stated user preferences, and exploring the viability of alternative approaches to designing recommender systems.

A. Harms Associated with Maximizing Predicted Engagement

In recent years, observers have linked online platforms to a range of harms. These include increased polarization and reductions in trust, direct incitement to violence up to and including acts of genocide, harms to mental and physical health, and harms stemming from privacy invasion, commercial surveillance, and user profiling, among others.³¹ While itemizing every potential harm is outside the scope of this report, it is useful to identify key vectors for harms to individuals, as defined by users themselves, that can stem from exposure to recommender systems. Building on prior work, these include:³²

- **Unwanted or harmful content:** Direct promotion by a recommender system that results in exposure to a harmful item, such as inducement to self-harm, bullying, or graphic violence.
- **Unwanted or harmful usage of the product:** Usage of a product that results in aggregate harmful effects on the user even if no particular item is objectionable, such as addiction,

³⁰ Stray et al., “What Are You Optimizing For?”

³¹ Bavel et al., “How Social Media Shapes Polarization”; Brailovskaia et al., “Experimental Longitudinal Evidence for Causal Role of Social Media Use and Physical Activity in COVID-19 Burden and Mental Health”; Park et al., “Global Mistrust in News”; United Nations Independent Investigative Mechanism for Myanmar, “Anti-Rohingya Hate Speech On Facebook”; Pasquale, *The Black Box Society*; Turov, *The Daily You*.

³² Lubin et al., “Social Media Harm Abatement.”

problematic or compulsive usage, reduction in sleep that impairs functioning, or reductions in markers of subjective well-being.

- **Unwanted or harmful contact:** While harmful contact (such as harassment or unwanted sexual advances) is not carried out directly via a recommender system, recommender systems can actively promote contacts,³³ groups, or networks that engage in or foster this behavior.³⁴
- **Unwanted or harmful usage of personal information:** While specific use of misappropriated user data or media (e.g., revenge porn and doxxing) may not be initiated by a recommender system, these items are sometimes promoted by recommender systems.

Concerns have also been raised about collective or systemic harms, risks, and impacts. Systemic risks are a central construct in the EU's DSA. The discourse that has ensued since its introduction and adoption into law has helped to clarify the concept.³⁵ Systemic harms tend to:

- be widespread or large-scale;
- result from interactions or feedback loops between the recommender system and multiple other entities (e.g., users, other platforms, or institutions);
- undermine important rights or institutions; and
- be difficult to reverse (although the impacts may be initiated by poor choices in the design of a recommender system, they have cascading effects that cannot be undone by merely improving the design of the system).

While academic research has identified connections between various harms and recommender systems designed to maximize predicted engagement, it faces important challenges, including lack of access to necessary data, inability to experiment with alternative designs in realistic settings, and legal risks associated with studying online platforms. Researchers have attempted to surmount these difficulties through various creative empirical methodologies, and though imperfect, findings point to specific associations between engagement and harms of different kinds.

One potential harm stems from the central role of recommender systems in extending social media use. Robust empirical literature has documented that among adolescents, extended use of social media (spurred by engagement-based designs) straightforwardly contributes to a decrease in time associated with healthier activities such as sleep.³⁶ When this happens, sleep may be disrupted

³³ Thiel et al., "Addressing the Distribution of Illicit Sexual Content by Minors Online."

³⁴ In several recent cases, plaintiffs have alleged harms relating to platform recommendation of minors' profiles to strangers with no friends in common or the recommendation of adult drug dealers' profiles to minors with no friends in common. See, e.g., "In Re: Social Media Adolescent Addiction Personal Injury Products Liability Litigation", District Court for the Northern District of California; "Neville v. Snap, Inc.", Superior Court of California.

³⁵ Sullivan and Pielmeier, "Unpacking 'Systemic Risk' Under the EU's Digital Service Act."

³⁶ Alonzo et al., "Interplay between Social Media Use, Sleep Quality, and Mental Health in Youth"; Brautsch et al., "Digital Media Use and Sleep in Late Adolescence and Young Adulthood"; Carter et al., "Association Between Portable Screen-Based Media Device Access or Use and Sleep Outcomes."

through various mechanisms, including delayed and worsened quality, increased psychological stimulation before bedtime, and distorted circadian rhythms from light emissions.³⁷ Indeed, research has found that adolescents often report using social media late at night and losing track of time when doing so.³⁸ This occurrence is concerning because insufficient sleep can affect various other health issues, such as the likelihood of learning problems, depression, and suicidal ideation.³⁹ Engagement-based feeds may plausibly contribute to these outcomes in adolescents, although more research is needed to examine this connection.

Engagement-based feeds may further have consequences for the user experience and behavioral norms of platforms. For instance, engagement may exacerbate negativity and polarization online. Though research attempting to link social media with political polarization in the aggregate has had mixed effects,⁴⁰ experimental studies have found that, compared with alternative designs, engagement-based ranking elevates negative emotions (including anger and sadness) and hostility toward outgroups among users, as well as the share of items expressing this negativity and hostility.⁴¹ Optimizing for engagement may also shape the kinds of items users are exposed to in detrimental ways. Empirical research has documented how engagement contributes to increased encounters with borderline abuse (such as insults and targeted cursing)⁴² and low-quality information about news events.⁴³ Each of these findings raises important concerns about how engagement directly affects users and their behavior, although further research and transparency are needed to adequately understand their consequences.⁴⁴

Notably, the guidance provided in this report is agnostic as to the type of underlying individual or systemic harms potentially mitigated should the recommendations be adopted. The guidance is targeted at upstream aspects of system design that could potentially mitigate a wide variety of harms and create a wide variety of benefits to users.

B. Harms to Minors

Research into the cognitive and social-emotional development of adolescents indicates they may be more vulnerable to risks associated with exposure to social media than adults.⁴⁵ Several traits unique

³⁷ LeBourgeois et al., “Digital Media and Sleep in Childhood and Adolescence.”

³⁸ Common Sense Media, “Constant Companion.”

³⁹ Paruthi et al., “Consensus Statement of the American Academy of Sleep Medicine on the Recommended Amount of Sleep for Healthy Children.”

⁴⁰ Kubin and von Sikorski, “The Role of (Social) Media in Political Polarization.”

⁴¹ Milli et al., “Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media”; Piccardi et al., “Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity.”

⁴² Bandy and Lazovich, “Exposure to Marginally Abusive Content on Twitter.”

⁴³ Moehring, “Personalization, Engagement, and Content Quality on Social Media.”

⁴⁴ These experimental studies exhibit important limitations, which range from potentially poor external validity to overreliance on user reports. One key issue is that empirical studies disproportionately focus on Twitter (now X). Historically this focus resulted from Twitter being a platform with easy-to-analyze text features and researcher-friendly API access, which has since been discontinued.

⁴⁵ See, e.g., Office of the Surgeon General, “Social Media and Youth Mental Health.”

to adolescent development affect their social media use.⁴⁶ Adolescence is often marked by a sensitivity to social acceptance from peers,⁴⁷ and adolescents' opinions and decisions may be more influenced by peers than by adults.⁴⁸ In addition, during adolescence, regions of the brain associated with emotional processing develop faster than those involved with reasoning and impulse control,⁴⁹ which may shape how adolescents use and respond to social media content.⁵⁰

In connection with these developmental realities, a broad body of research has demonstrated that children and teens experience negative consequences from social media use in some cases,⁵¹ and that they are often subject to engagement-maximizing tactics.⁵² By comparison, not enough research has decomposed how these harms result from specific elements of platform design, such as optimization of recommender systems, although a small literature base has identified mechanisms by which recommender system design undermines child well-being.

First, some studies have examined the prevalence of self-harm, violence, and other content categories on video-sharing platforms popular among children, such as YouTube and Instagram, and observed that recommendations often feature this content.⁵³ Second, a series of experimental studies in which researchers opened accounts purporting to be minors has found that sustained engagement with harmful content substantially increases the rates at which this content is recommended, raising concerns about the interaction between recommender systems and vulnerable minors (such as those with severe mental health issues).⁵⁴ Finally, a handful of studies have warned about the potential for recommender systems to connect children's accounts with sexual predators and other dangerous individuals.⁵⁵

C. User Preferences and Satisfaction

Beyond the specific harms to adults and youth identified above, one major theme of research on recommender systems is that optimizing for predicted engagement may lead to recommendations that are not always aligned with user satisfaction.

⁴⁶ Crone and Konijn, "Media Use and Brain Development during Adolescence."

⁴⁷ Somerville, "The Teenage Brain."

⁴⁸ Ibid.

⁴⁹ Casey et al., "The Adolescent Brain."

⁵⁰ 5Rights Foundation, "Pathways"; Costello et al., "Algorithms, Addiction, and Adolescent Mental Health"; Chen et al., "The Engagement-Prolonging Designs Teens Encounter on Very Large Online Platforms"; Pizzo Frey et al., "Recommendation Systems in Social Media."

⁵¹ National Academies of Sciences, Engineering, and Medicine, *Social Media and Adolescent Health*.

⁵² Costello et al., "Algorithms, Addiction, and Adolescent Mental Health."

⁵³ Bryant, "Instagram Actively Helping Spread of Self-Harm among Teenagers, Study Finds"; Radesky et al., "Algorithmic Content Recommendations on a Video-Sharing Platform Used by Children"; Papadamou et al., "Disturbed YouTube for Kids."

⁵⁴ Amnesty International, "Driven into Darkness"; Center for Countering Digital Hate, "Deadly by Design"; Hilbert et al., "#BigTech @Minors."

⁵⁵ Pizzo Frey et al., "Recommendation Systems in Social Media"; Thiel et al., "Addressing the Distribution of Illicit Sexual Content by Minors Online."

Optimizing for predicted engagement can lead to negative experiences with individual recommendations, dissatisfaction with the overall amount of time spent or experience using the product, or both.⁵⁶ Today it is common for platforms to rely on engagement signals as proxies for users' desires,⁵⁷ implying that their behaviors reveal accurate information about their preferences.⁵⁸ Under this logic, training recommender systems on behavioral data and designing them to make predictions about user behavior should result in recommendations that satisfy users and align with their preferences.⁵⁹

However, research has exposed how that assumption is not always valid, especially given insights into the mechanics of individual decisionmaking.⁶⁰ This is often formalized as the “dual systems” model of understanding choices, wherein an individual has one impulsive, mindless, or myopic self (sometimes referred to as System 1) and one forward-looking, thoughtful self (System 2) looking out for the individual over the long term.⁶¹ This can result in an individual's actual, underlying preferences being short-circuited by their impulsive preferences. The individual's impulsive behaviors will not necessarily represent their long-term interests. As a result, recommended content, though successful in inducing engagement, may fail to satisfy users and steer them toward behaviors they later regret, such as staying up late at night.⁶²

To illustrate how this manifests in the design of recommender systems, one study analogizes consuming content to eating potato chips at a party.⁶³ A person attending a party might eat a whole bowl of potato chips, and the party host might take this as a sign to refill the chip bowl. But perhaps the party guest is eating impulsively, when in fact they have a long-term goal to be eating healthier food. The guest's impulsive behavior is misaligned with their underlying preference, but the host interprets the behavior as a sign of what the guest wants.

In a similar manner, impulsively dwelling on, clicking, and liking certain content (say, content that is risky in some way) does not necessarily reflect the user's forward-looking desires to the platform. A platform that concludes from this engagement that the user must want more and more of this content would be falsely assuming that the user's impulsive, mindless, or myopic behavior equates with their long-term preferences.

⁵⁶ See, e.g., Allcott et al., “Digital Addiction”; Cho et al., “Reflect, Not Regret.”

⁵⁷ The scope of information provided by these signals matters significantly for the quality of recommendations. For example, when a recommender system designed to rank news items lacks enough personal data to inform its recommendations, its performance is worse relative to human curation. See Peukert et al., “The Editor and the Algorithm.”

⁵⁸ Kleinberg et al., “The Challenge of Understanding What Users Want.”

⁵⁹ Ibid.

⁶⁰ Agan et al., “Automating Automaticity”; Agarwal et al., “System-2 Recommenders”; Christakopoulou et al., “Deconfounding User Satisfaction Estimation”; Christakopoulou et al., “Reward Shaping for User Satisfaction”; Kleinberg et al., “The Challenge of Understanding What Users Want”; Milli et al., “From Optimizing Engagement to Measuring Value”; Stray et al., “What Are You Optimizing For?”

⁶¹ Evans and Frankish, *In Two Minds*; Samson and Voyer, “Two Minds, Three Ways.”

⁶² Common Sense Media, “Constant Companion.”

⁶³ Kleinberg et al., “The Challenge of Understanding What Users Want.”

This dynamic often occurs on large online platforms, likely having consequences for both the well-being of users (as recommendations cause them to behave in ways they later regret) and the general tenor of behaviors promoted. For example, one study identified that engagement-based designs strongly incentivize creators to compete for attention, such as by using sexual, racial, or humor-based standard tactics to shape the content they offer.⁶⁴ Another study has demonstrated that, when users engage with items, they often do so impulsively. Training recommender systems on this data about these impulsive behaviors and heuristics can bias their recommendations toward perpetuating that type of behavior.⁶⁵

D. Alternatives to Maximizing Predicted Engagement

1. Chronological and Non-Personalized Feeds

In both the US and the EU, laws and policies promoting chronological feeds or limiting the use of personal data to customize feeds have been proposed and adopted as alternatives to designs that maximize predicted engagement. Although these two approaches are distinct, chronological feeds are sometimes referenced as a simple approach to comply with requirements intended to limit the use of personal data in feed design. The appeal of both options as policy approaches is clear: they are simple to understand conceptually, and, in the case of chronological feeds, they have been a commonly-deployed design since the invention of online communications services (including in email, message boards, and social media itself). But existing evidence reveals the drawbacks of these approaches.

Research shows that chronological feeds can decrease engagement and cause users to switch back to engagement-optimized feeds (if available).⁶⁶ Chronological feeds also have multifaceted effects on user experience that are not necessarily positive. One major finding is that chronological feeds can shift the mix of recommended items in unexpected ways: these feeds may increase a user's relative exposure to abuse, decrease content from accounts in their social network, and amplify the prevalence of political and untrustworthy content.⁶⁷ Moreover, chronological feeds may create a recency bias that incentivizes "spammy" posting behavior,⁶⁸ and they are not workable for all types of platforms (e.g., streaming services like Netflix and Spotify). This makes chronological feeds a suboptimal choice in many cases if the goal is to deliver high-quality experiences to the user.

⁶⁴ Common Sense Media, "Who Is the 'You' in YouTube?"

⁶⁵ Agan et al., "Automating Automaticity."

⁶⁶ Bandy and Lazovich, "Exposure to Marginally Abusive Content on Twitter"; Guess et al., "How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?"; Moehring, "Personalization, Engagement, and Content Quality on Social Media."

⁶⁷ Bandy and Lazovich, "Exposure to Marginally Abusive Content on Twitter"; Guess et al., "How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?"

⁶⁸ Bengani, "What's Right and What's Wrong with Optimizing for Engagement."

Suggesting that users must choose between today's engagement-optimized feeds and chronological or non-personalized feeds creates a false choice. In reality, the design space for recommender system optimization is vast. The next section explains other design approaches.

2. Better Approaches

Noting the deficiencies with both chronological feeds and recommender systems designed to maximize predicted engagement, researchers have explored how these systems could be redesigned to explicitly further different values. Theoretical studies have demonstrated that recommender systems need not be designed to maximize predicted engagement and that it is possible to balance engagement or even replace it with more prosocial values.⁶⁹ For example, one study enumerated more than 30 possible values such as self-expression, informativeness, and safety, that could inform the design of recommender systems instead of engagement.⁷⁰

An abundance of research has experimented with implementing designs for recommender systems optimized for some objective other than maximizing predicted engagement. These approaches tend to be optimized around one or more of the following measures: bridging, survey responses, or quality.

Bridging

Social media platforms frequently become battlegrounds for conflict between different social groups. In response, researchers have proposed redesigning recommender systems to foster mutual trust and understanding across these social divides – an approach known as "bridging."⁷¹ Rather than attempting to eliminate conflict entirely, bridging aims to transform it into something more constructive.⁷² This can be achieved through algorithmic recommendations that prioritize items promoting productive dialogue or positive emotions.

This approach to designing recommender systems is quite different from optimizing for predicted engagement. For example, scholars have demonstrated that items can be ranked according to the approval they obtain from diverse users or whether they receive radically different kinds of engagement from different social groups.⁷³ Thus far, bridging-based recommender systems have seen only limited real-world deployments. Several social media platforms have incorporated "Community Notes" features, which use bridging-based systems to select user-generated notes to display that provide context about content posted by other users.⁷⁴

⁶⁹ Agarwal et al., "System-2 Recommenders"; Besbes et al., "The Fault in Our Recommendations"; Milli et al., "From Optimizing Engagement to Measuring Value"; Singh et al., "Building Healthy Recommendation Sequences for Everyone"; Stray et al., "Building Human Values into Recommender Systems."

⁷⁰ Stray et al., "Building Human Values into Recommender Systems."

⁷¹ Ovadya and Thorburn, "Bridging Systems."

⁷² Stray, "Designing Recommender Systems to Depolarize."

⁷³ Ovadya and Thorburn, "Bridging Systems."

⁷⁴ See, e.g., The YouTube Team, "Testing New Ways to Offer Viewers More Context and Information on Videos"; Wirtschafter and Majumder, "Future Challenges for Online, Crowdsourced Content Moderation"; Wojcik et al., "Birdwatch."

Surveys

Currently, it is common practice for social media platforms to survey users about their feeds.⁷⁵ This usually occurs in two forms: questions about specific items (e.g. Facebook’s “Is this content good for the world?” survey) and more general questions about a user’s subjective experience (e.g., YouTube’s “Rate your YouTube experience today” survey).⁷⁶ Platforms generally make more use of engagement data than survey data when recommending content to users because the former is much cheaper to obtain in large amounts.⁷⁷ Survey data is thus less available and may exhibit bias that must be taken into account. To overcome these challenges, platforms use the survey data they collect from a fraction of users about a fraction of items to predict how the rest of the user population might feel about similar items.⁷⁸ Figure 3 shows example surveys.

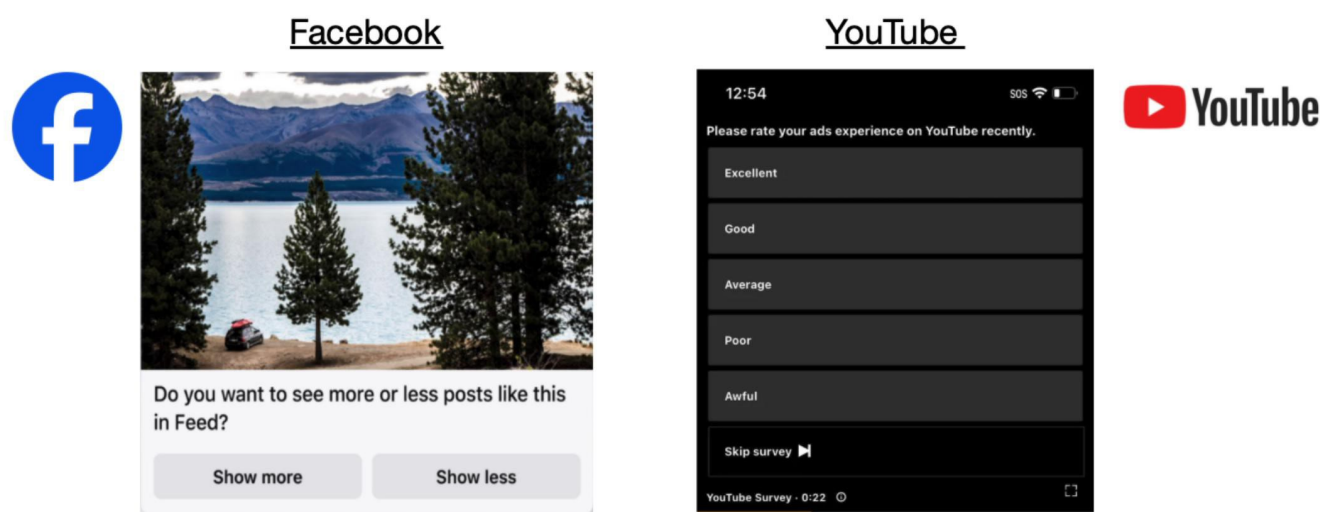


Figure 3. Social media platforms often run surveys on a small subset of their users. Above are two examples from Facebook and YouTube featuring surveys asking users whether they want to see similar content and about the quality of ads.

Many scholars have proposed expanding the role surveys play in content recommendation.⁷⁹ For example, bridging-based recommender systems could make use of surveys by directly asking users how a specific item or their feeds as a whole affect their perceptions of political outgroups.⁸⁰ Recommender systems designed in this manner could assign rank scores based on predictions about how users would respond to survey questions. Using surveys in this way may help platforms detect unwanted experiences and surface higher-quality content.⁸¹

⁷⁵ Cunningham et al., “What We Know About Using Non-Engagement Signals.”

⁷⁶ Ibid.

⁷⁷ Christakopoulou et al., “Reward Shaping for User Satisfaction.”

⁷⁸ Ibid.; Cunningham et al., “What We Know About Using Non-Engagement Signals.”

⁷⁹ Fast et al., “Unveiling the Neely Ethics & Technology Indices”; Iyer et al., “How User Experience Metrics Complement ‘Content That Requires Enforcement’”; Milli et al., “Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media”; Stray, “Designing Recommender Systems to Depolarize”; Stray, “Dependent Variables.”

⁸⁰ Stray, “Designing Recommender Systems to Depolarize.”

⁸¹ Cunningham et al., “What We Know About Using Non-Engagement Signals”, 10-11.

Quality

Researchers have proposed designing recommender systems to rank items based on some specified dimension of quality.⁸² For example, a quality-based recommender system could analyze the features of an item's content for whether it uses profanity and other toxic language (such as slurs and targeted insults) and then assign rank scores to items based on this analysis.

This example illustrates how recommender systems can be explicitly designed to further a specific quality objective. Many systems developed along these lines have demonstrated the viability of this approach to ranking.⁸³ Their implementation at scale may be challenging because relevant stakeholders do not always agree about whether or which values should be explicitly promoted through design.⁸⁴

3. Implications for Recommender System Design

Empirical research has shown that recommender systems designed to promote values other than engagement are both viable and successful at promoting user satisfaction. A number of studies have experimentally tested the effects of reranking recommended items using browser extensions on real platforms.⁸⁵ These experiments find that users notice when changes are made in how items are recommended; that user experiences change in response to differences in recommended items; and that users are able to make effective use of tools that strengthen their control.⁸⁶ Additional empirical research has also concluded that implementing alternative values in recommender systems is possible with only low to moderate trade-offs in measured engagement.⁸⁷

At a lower level in recommender system design, recent research has demonstrated that the signals employed by recommender systems can affect the characteristics of items recommended as well as the viability of using non-engagement based signals at scale.⁸⁸ When recommender systems are trained strictly on the most readily available engagement signals (clicks, dwell time, etc.), they will tend to learn to choose items that drive those forms of engagement.⁸⁹ Conversely, balancing different kinds

⁸² Singh et al., "Building Healthy Recommendation Sequences for Everyone"; Moehring, "Personalization, Engagement, and Content Quality on Social Media"; Stray et al., "Building Human Values into Recommender Systems."

⁸³ See, e.g., Jigsaw, "Perspective API - How It Works"; NewsGuard, "News Reliability Ratings."

⁸⁴ Stray et al., "Building Human Values into Recommender Systems."

⁸⁵ Piccardi et al., "Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity."

⁸⁶ Ibid.

⁸⁷ Moehring, "Personalization, Engagement, and Content Quality on Social Media"; Piccardi et al., "Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity."

⁸⁸ Agan et al., "Automating Automaticity"; Christakopoulou et al., "Deconfounding User Satisfaction Estimation"; Christakopoulou et al., "Reward Shaping for User Satisfaction," Kleinberg et al., "The Challenge of Understanding What Users Want"; Peukert et al., "The Editor and the Algorithm."

⁸⁹ Christakopoulou et al., "Reward Shaping for User Satisfaction," 4.

of signals can increase reported user satisfaction. For example, researchers have shown that combining survey-based signals with engagement-based signals achieves this improvement.⁹⁰

This work also points to the importance of allowing personalization of recommender systems, even (or especially) if they are not optimized for predicted engagement. Personalization has independent benefits that users may value. Research has demonstrated that personalization increases user satisfaction, improves the quality of content delivered, and decreases the search costs of finding high-quality content.⁹¹ Not all uses of engagement signals and predictions are personalized, and it is possible to personalize recommender systems without optimizing for predicted engagement. For example, personalized systems can be designed to assign higher scores to items in a user's preferred language, from a particular geographic area, or consistent with their stated preferences. Personalization can also be key to recommender systems that rely on individuals' data to enhance bridging, that leverage individuals' survey responses or predicted survey responses, or that customize quality-based ranking to individual users, for example.

Public policy that seeks to address engagement-based harms by eliminating personalization therefore aims at the wrong target. Approaches that directly address the maximization of predicted engagement are preferable, especially since personalization can strengthen the viability of non-engagement-based ranking strategies.⁹²

As a whole, these findings indicate that the universe of possible designs for recommender systems does not consist of a binary choice between optimizing for predicted engagement and chronological feeds. Although technical barriers present some limitations to deployment, the core reason that these alternative implementations are largely absent from the marketplace is misaligned incentives between the designers of recommender systems and the interests of their users and society at large. Public policy can support the deployment of alternative designs for recommender systems. Policymakers have already begun this work.

⁹⁰ Christakopoulou et al., “Deconfounding User Satisfaction Estimation”; Christakopoulou et al., “Reward Shaping for User Satisfaction.”

⁹¹ Budzinski and Lindstädt-Dreusicke, “Data (r)Evolution”; Donnelly et al., “Welfare Effects of Personalized Rankings”; Rafieian and Yoganarasimhan, “AI and Personalization.”

⁹² Moehring, “Personalization, Engagement, and Content Quality on Social Media.”

IV. Policy Landscape

In the US, recent federal and state approaches to regulation of recommender systems and algorithmic amplification have focused on a few key areas: limiting Section 230 immunity from liability, enshrining transparency requirements into law, and consumer protection provisions such as prohibiting algorithmic misuse and discrimination. Few of these bills have passed into law, demonstrating the challenges of regulation in this sphere. Constitutional issues play a role as well, in particular the First Amendment, making these concerns the subject of several litigation efforts.

The EU's approach marks a contrast with nascent US regulatory efforts, with the DSA establishing transparency, choice, and risk assessment requirements with respect to platforms' recommender systems.

Federal Legislation

The 117th and 118th Congresses saw several bills aimed at algorithmic amplification and recommender systems. Bills introduced that sought to limit Section 230 immunity from liability for social media companies based upon their amplification or recommendation of certain content include the Platform Integrity Act, which would have limited the application of the "Good Samaritan" provision of Section 230, regarding the publication of third party content, if a provider or user "promoted, suggested, amplified, or otherwise recommended" the content;⁹³ the Justice Against Malicious Algorithms Act (focused on content causing physical or emotional injury);⁹⁴ the Protecting Americans from Dangerous Algorithms Act (focused on content relating to interference with civil rights or content concerning acts of international terrorism);⁹⁵ the DISCOURSE Act (focused on content provided to a user who neither requested nor searched for it);⁹⁶ and the Health Misinformation Act (focused on content concerning health misinformation during a declared public health emergency).⁹⁷

Other efforts included transparency-based bills such as the Algorithmic Justice and Online Platform Transparency Act, which set forth reporting and record keeping requirements related to the algorithmic use of personal information;⁹⁸ the Algorithmic Accountability Act, which sought FTC-led impact assessments of automated decision systems on critical issues that have a significant effect on consumers;⁹⁹ the Platform Accountability and Transparency Act, which would require platforms to provide researcher access to platform data and proactively release certain information to researchers and the public, including information concerning ranking and recommendation algorithms;¹⁰⁰ and the

⁹³ United States, "Platform Integrity Act."

⁹⁴ United States, "Justice Against Malicious Algorithms Act."

⁹⁵ United States, "Protecting Americans from Dangerous Algorithms Act."

⁹⁶ United States, "DISCOURSE Act."

⁹⁷ United States, "Health Misinformation Act."

⁹⁸ United States, "Algorithmic Justice and Online Platform Transparency Act."

⁹⁹ United States, "Algorithmic Accountability Act."

¹⁰⁰ United States, "Platform Accountability and Transparency Act."

Kids Online Safety and Privacy Act (KOSPA), which would require platforms to disclose information concerning their use of personalized recommendation systems, among other things.¹⁰¹

Taken together, these bills indicate a recognition by legislators of the unknowns wrought by recommender systems and the resulting need for platform transparency. They also reveal a lack of consensus as to the means to achieve accountability for algorithmic recommendations.¹⁰² Despite interest from both major political parties, none of these bills have become law.

State Legislation

State-level legislative efforts in algorithmic regulation point to lawmakers' concerns about the impact of recommender systems on minors and on the suppression of certain views (sometimes referred to as platform "censorship" or "shadow banning").

Bills targeting addictive feeds – specifically, the use of algorithms to spur online engagement by minors – have been proposed in several states. Those that were passed into law include California's Protecting Our Kids from Social Media Addiction Act¹⁰³ and New York's Stop Addictive Feeds Exploitation (SAFE) for Kids Act.¹⁰⁴ These statutes seek to protect children by limiting the personal data that can form the basis of algorithmic recommendations. North Carolina's Social Media Algorithmic Control in IT Act, which prohibits the use of minors' data for algorithmic recommendations, is pending.¹⁰⁵ Other state proposals have emphasized offering minors alternatives to algorithmic feeds, requiring platforms to provide non-personalized or chronological feeds as options.¹⁰⁶

Other bills have sought to address algorithmic ranking through transparency requirements rather than direct restrictions. Minnesota's Prohibiting Social Media Manipulation Act was passed into law in 2024 and requires platforms to disclose whether and how they assess content quality and explicit user preferences and how those signals are weighted in algorithmic systems in relation to engagement signals.¹⁰⁷ The law also requires disclosure of all product experiments conducted on more than 1,000 users including any negative effects resulting from those experiments.¹⁰⁸ California's transparency law, AB 587 (enacted in 2022 and subsequently narrowed following litigation) takes a different approach, focusing on mandatory disclosures of content moderation practices.¹⁰⁹

Age-appropriate design codes have been another common approach to broader concerns about youth online safety, with many state bills modeled after California's Age-Appropriate Design Code

¹⁰¹ United States, "Kids Online Safety and Privacy Act."

¹⁰² For in-depth exploration of some of the legal complexities, see Austin and Levy, "Speech Certainty"; Balkin, "Free Speech Versus the First Amendment."

¹⁰³ California, "Protecting Our Kids from Social Media Addiction Act."

¹⁰⁴ New York, "Stop Addictive Feeds Exploitation (SAFE) for Kids Act."

¹⁰⁵ North Carolina, "Social Media Algorithmic Control in IT Act."

¹⁰⁶ Colorado, "SB158."

¹⁰⁷ Minnesota, "Prohibiting Social Media Manipulation Act."

¹⁰⁸ Ibid.

¹⁰⁹ California, "AB 587."

Act,¹¹⁰ including Maryland's Kids Code¹¹¹ that was enacted into law in 2024, and the 2024 Vermont Kids Code that passed as part of a broader consumer privacy bill but was then vetoed.¹¹² These bills generally require platforms to assess the harms to minors stemming from algorithmic recommendations and prohibit the use of deceptive patterns in product design that manipulate minors into taking actions against their own interest. Other states have focused on banning or restricting minors' use of social media altogether or during certain periods of the day, prohibiting extended-use designs (e.g., autoplay, infinite scroll, push notifications, gamification), requiring age assurance and parental controls, or prohibiting algorithms from recommending harmful content to minors.¹¹³

Several states, led initially by Florida and Texas, have proposed or enacted bills aimed at combating so-called "censorship," "shadow banning" and "post-prioritization" of specific content or user accounts.¹¹⁴ These bills would affect algorithmic ranking by prohibiting platforms from using algorithms to "disfavor" certain content. This approach would also require platforms to disclose how their algorithms work and allow users to opt out in favor of chronological feeds.¹¹⁵

Legal Challenges to State Laws

Several of the state laws that have passed have been challenged in court, largely on First Amendment grounds. Many of these laws have been enjoined while litigation continues.

Challenges to laws that impose broad design obligations,¹¹⁶ transparency mandates,¹¹⁷ and social media bans or age assurance requirements¹¹⁸ have largely succeeded thus far in halting enforcement while litigation is ongoing. For example, in 2024 the Ninth Circuit Court of Appeals affirmed a preliminary injunction against the California Age-Appropriate Design Code Act's requirement for businesses to assess and mitigate potential harm to children, citing First Amendment violations, while vacating and remanding other provisions for further consideration.¹¹⁹ In *X Corp. v. Bonta*, the Ninth Circuit Court of Appeals reversed the district court's denial of a preliminary injunction, finding that California's AB 587 likely violates the First Amendment by compelling social media companies to disclose their content moderation policies regarding specific categories such as hate speech and misinformation.¹²⁰ A string of district court judgments have prevented broadly scoped online child

¹¹⁰ California, "California Age-Appropriate Design Code Act."

¹¹¹ Maryland, "Maryland Kids Code."

¹¹² Vermont, S 289."

¹¹³ See, e.g., Texas, "SCOPE Act"; Utah, "SB 194 Social Media Regulation Amendments"; Virginia, "Consumer Data Protection Act."

¹¹⁴ Florida, "SB 7072"; Texas, "HB 20."

¹¹⁵ See, e.g., Hawaii, "Anti-Big-Tech Censorship Act"; Minnesota, "SF 2716."

¹¹⁶ United States District Court for the Northern District of California, "NetChoice v. Bonta."

¹¹⁷ United States Court of Appeals for the Ninth Circuit, "X Corp. v. Bonta."

¹¹⁸ United States District Court for the District of Utah, "NetChoice v. Reyes"; United States District Court for the Western District of Texas, "Computer & Communications Industry Association and NetChoice v. Paxton."

¹¹⁹ United States Court of Appeals for the Ninth Circuit, "X Corp. v. Bonta."

¹²⁰ Ibid.

safety laws from going into effect in Arkansas, Ohio, Mississippi, Texas, and Utah on the basis of potential First Amendment violations.¹²¹

Similarly broad laws passed in Florida and Texas aimed at social media content moderation and promotion practices are at issue in *NetChoice v. Moody* and *NetChoice v. Paxton*. Here, the Eleventh and Fifth Circuits, respectively, came to differing conclusions as to the constitutionality of the subject laws after federal district courts in both Florida and Texas enjoined the laws, noting the likely success of the trade association plaintiffs on First Amendment grounds.¹²² The Supreme Court vacated the decisions and remanded for further proceedings due to the lower courts' failure to conduct an analysis of the laws' applications in the First Amendment context.¹²³ In so doing, the Court left open the possibility that some platform regulation—including, possibly, regulation concerning algorithmic amplification where algorithms responded “solely to how users act online”—could be compatible with the First Amendment.¹²⁴

In one of the first court opinions to consider more narrowly tailored approaches to regulating algorithmic feeds, the judge similarly concluded that such regulation may be consistent with the First Amendment. Ruling on a challenge to California's Protecting Our Kids from Social Media Addiction Act in late 2024, the District Court for the Northern District of California allowed the law's provisions restricting algorithmic feeds for minors to go into effect, while enjoining other provisions.¹²⁵ Citing *Netchoice v. Moody*, the decision distinguished between algorithmic components that convey a point of view, such as content moderation decisions, and those that are functional, such as optimization for time spent. This distinction may open the door to further algorithmic regulation within the US.

Section 230 Jurisprudence

In early cases, judges granted Section 230 immunity to defendants against claims that using a recommender system made a platform the co-creator of harmful content. For example, in *Force v. Facebook*, the Second Circuit held that Facebook's friend and content recommendations did not make the company the co-creator of terrorist recruiting content.¹²⁶ Similarly, in *Dyroff v. Ultimate Software Group*, the Ninth Circuit ruled that a recommender system did not make a website the co-creator of messages advertising illegal drug sales.¹²⁷ These courts did not wholesale immunize recommender systems under Section 230, but clarified that plaintiffs could not evade the law's reach by alleging that recommendation algorithms made websites the authors of harmful content.

However, some courts have identified limits to Section 230's coverage of algorithmic systems. In *Lemmon v. Snap*, the Ninth Circuit held that Section 230 did not bar claims focused on Snapchat's

¹²¹ See, e.g., United States District Court for the District of Utah, “NetChoice v. Reyes.”

¹²² Supreme Court of the United States, “NetChoice v. Moody.”

¹²³ *Ibid.*, 12.

¹²⁴ *Ibid.*, 22.

¹²⁵ United States District Court for the Northern District of California, “NetChoice v. Bonta.”

¹²⁶ United States Court of Appeals for the Second Circuit, “Force v. Facebook.”

¹²⁷ United States Court of Appeals for the Ninth Circuit, “Dyroff v. The Ultimate Software Group.”

"Speed Filter" feature design, distinguishing between content-focused moderation (protected) and dangerous product features (not protected).¹²⁸ More recently, the Third Circuit ruled in *Anderson v. TikTok* that Section 230 did not immunize TikTok for recommending a dangerous challenge video which allegedly led to the death of a minor.¹²⁹

Attorneys General have filed joint and individual suits alleging deceptive design and harmful practices, including a 42-state lawsuit alleging Meta's social media products are harmful to youth.¹³⁰ Private plaintiffs have likewise filed hundreds of lawsuits on behalf of youth, families, and school districts based on product liability, negligence, misrepresentation, deception, and a variety of other claims, some of which are based on harms that plaintiffs connect to content or account/friend recommendations. Many of these cases have been consolidated in federal and state courts and litigation is ongoing.¹³¹

The Supreme Court has yet to address the scope of Section 230 immunity for content recommendations, sidestepping the issue in two 2023 decisions, *Gonzalez v. Google* and *Taamneh v. Twitter*. In *Gonzalez*, families of those killed in ISIS terrorist attacks sued Google, alleging that it had aided and abetted those attacks by allowing ISIS to post videos to YouTube and by recommending those videos to users algorithmically. In *Taamneh*, relatives of an ISIS attack victim in Turkey sued Twitter, alleging that as ISIS had used Twitter to expand its reach and Twitter both knew ISIS had done so and failed to take appropriate countermeasures, Twitter had aided and abetted an act of international terrorism. The Supreme Court held that Twitter could not be held liable for aiding and abetting a specific attack when they had not knowingly assisted in it.¹³² The Court determined that *Gonzalez* could be addressed on similar grounds, and declined to take up the question of the scope of Section 230.¹³³

EU Regulation

The EU DSA establishes transparency, accountability, and user control obligations related to algorithmic recommender systems. Platforms must disclose the main parameters of their algorithms.¹³⁴ The DSA requires very large online platforms ("VLOPs," with over 45 million users in the EU) to provide users with at least one recommender system option that is not based on "profiling."¹³⁵ The platforms must make this alternative system easily accessible in their interface. Additionally, platforms must implement mechanisms to mitigate systemic risks, such as the spread of disinformation or harm to

¹²⁸ United States Court of Appeals for the Ninth Circuit, "Lemmon v. Snap."

¹²⁹ United States Court of Appeals for the Third Circuit, "Anderson v. TikTok."

¹³⁰ See Feiner, "Meta Sued by 42 Attorneys General Alleging Facebook, Instagram Features Are Addictive and Target Kids."

¹³¹ See, e.g., United States District Court for the Northern District of California, "In Re: Social Media Adolescent Addiction/Personal Injury Products Liability Litigation."

¹³² Supreme Court of the United States, "Taamneh et al. v. Twitter."

¹³³ Supreme Court of the United States, "Gonzalez et al. v. Google," 2. ("[I]t has become clear that plaintiffs' complaint—independent of §230—states little if any claim for relief. As plaintiffs concede, the allegations underlying their secondary-liability claims are materially identical to those at issue in Twitter.")

¹³⁴ European Union, "Digital Services Act," Article 27.

¹³⁵ European Union, "Digital Services Act," Article 38.

vulnerable populations, arising from their recommendation algorithms. For VLOPs, the DSA mandates independent audits of algorithmic systems to ensure compliance and assess their societal impacts.

The European Commission has begun its DSA enforcement, sending preliminary requests for information and opening multiple investigations into different platforms to examine non-compliance with a variety of DSA provisions. The Commission has launched several inquiries into different platforms related to recommender system provisions, all of which are pending resolutions.¹³⁶

V. Core Policy Guidance

Current approaches to regulating recommender systems have made important attempts at mitigating harms associated with engagement-based designs, yet there is room for improvement. Some approaches have thus far focused on broad restrictions – banning or restricting personalization, algorithmic feeds, or both – while others are more nuanced. Chronological feeds are often promoted in policy discussions as the simple solution to algorithmic concerns. However, as the analysis in Section III shows, chronological feeds have drawbacks, and there are numerous other ways to design systems that provide users with valuable, high-quality experiences. Effective regulation should encourage the development of these alternative approaches, which can address potential harms while preserving the benefits that thoughtful use of engagement data and personalization can offer.

This section provides policy guidance designed with the US legal framework in mind. While there is no guarantee that the constitutionality of legislative or regulatory efforts based on these guidelines would be upheld, this guidance attends to concerns about the potential for regulation to implicate speech rights under the First Amendment and platform liability immunity under Section 230. These are policy guidelines only; developing legislative text to support any of the guidelines would require nuance based on evolving case law.

Section VI below offers additional policy guidance that may be more feasible to implement outside the US where different legal frameworks govern corporate and individual speech and liability. Some of these proposals may be implementable in the US, but they are crafted with non-US jurisdictions in mind.

Both sets of guidance can be applied across common features powered by recommender systems, including news feeds/timelines, ads, account/group/channel recommendations, notifications, and more. They can also apply to companies of different sizes and to a variety of different services where recommender systems are in use, including social media, search engines, streaming, e-commerce, and gaming. The task of defining which specific entities are covered is left for those who may adopt this guidance in law or policy.

¹³⁶ European Commission, “Supervision of the Designated Very Large Online Platforms and Search Engines under DSA”; Husovec, “The DSA Newsletter #6.”

Long-Term Value to Users

The policy guidance below makes frequent reference to long-term value to users. This concept encapsulates the objective of aligning recommender system design to outcomes that prioritize users' deliberative, forward-looking aspirations or preferences – in the “System 2” thinking sense – while remaining generic enough to allow different platforms to design their own way of achieving it.

For example, platforms that are optimized to support long-term value to users:

- may ask users directly to state their explicit preferences;
- may rely on surveys, quality indicators selected by the user, or predictions of each;
- may rely on signals that are deliberative, clear, or onerous; or
- may combine aspects of these or other approaches.

On many platforms, the most accurate data about long-term user value comes from users directly expressing preferences through user controls and settings. This should be viewed as the most robust approach to understanding long-term user value, where this data is available.

However, user controls are typically only adopted by a fraction of users, and for some recommender systems it may not be workable for platforms to offer user controls. In the absence of any more explicit preference data from users, the list of accounts that users choose to follow or subscribe to (on platforms that support such functionality) could be considered an indicator of long-term preference. Some platforms survey users to identify long-term value, and while survey populations can be representative of the user base, they typically engage a small subset of users.

If explicit preferences or survey responses are not available, approaches that predict or extrapolate from user preference or survey data to the rest of the user population may be a reasonable alternative. Unlike optimizing for short-term predicted engagement, these approaches are based on explicit information supplied by users, albeit not all users. Platforms may also consider adopting deliberative processes where small subsets of users engage in intensive processes to identify how recommender systems can be optimized for long-term value, and applying those results across the user base.¹³⁷

One way of knowing when a platform's design fails to serve its users' long-term preferences or aspirations is when meaningful numbers of users regret their experiences on the platform or report

¹³⁷ See, e.g., Huang et al., “Collective Constitutional AI” and Ovadya, “Towards Platform Democracy” for discussions of deeply deliberative preference elicitation in this context.

a loss of self-control. External surveys of user regret¹³⁸ can serve as a means to validate that platforms are indeed optimizing for long-term preferences or aspirations. These outcomes indicate design patterns that undercut what users want from their experience on a platform over the long run.

Optimizing recommender systems to maximize predicted short-term engagement does *not* typically promote long-term value. The mere fact that recommended items succeed in engaging users is not sufficient to establish that these recommendations align with long-term value to users. Rather, predictions of long-term value must be supported by evidence of explicit, expressed desires held by individual users or representative subsets of users, not ambiguous behaviors that may correlate with inferred “preferences.”

A. Design Transparency

Limited public disclosures about how recommender systems work have existed for years. What began with a few companies voluntarily disclosing aspects of their system design has led to co-regulatory and regulatory frameworks that now mandate disclosures in some jurisdictions, most notably for platforms in the EU as a result of the DSA.

Article 27 of the DSA requires designated platforms to disclose information about the design of their recommender systems in their terms of service.¹³⁹ Existing disclosures focus on what signals are processed and what predictions are made by each recommender system, and essentially nothing about the metrics used to evaluate their design.

While these disclosures go into detail about the signals used, few give information about which signals tend to be weighted more heavily, instead merely listing which signals play any role at all.

Unsurprisingly, all disclosures state that recommender systems process information about a user’s engagement (e.g., how many posts of a certain type they have liked). The two platforms that disclose any information about how signals are weighted – TikTok and Snap – also share that engagement signals receive the most weight.¹⁴⁰

Less information is provided about predictions, and none at all is disclosed about metrics. The disclosures establish that all platforms make predictions about how likely users are to engage with recommended items, but they provide no information about the relative weights of each prediction.

¹³⁸ For example, in 2024, the Harris Poll reportedly found that almost half of Generation Z users regret the invention of TikTok, X, and Snapchat, while far fewer regret Youtube or Netflix. See Skiera, “What Gen Z Thinks about Its Social Media and Smartphone Usage.”

¹³⁹ European Union, “Article 27, Recommender System Transparency.”

¹⁴⁰ Snap, “How We Rank Content on Discover”; Snap, “How We Rank Content on Spotlight”; TikTok, “How TikTok Recommends Content.”

And none of the platforms disclose information about which metrics are used to evaluate the success of recommender systems and the teams who designed them.

The mandatory disclosure of information about weights and metrics would allow outside experts, regulators, and the public to understand the tradeoffs being made in the design of recommender systems. These disclosures would allow for comparison of designs across different systems and over time. And the requirement to disclose would motivate platforms to optimize their designs and internal incentive structures in ways that demonstrate their attentiveness to long-term user value and satisfaction.

Guidelines:

Platforms must publicly disclose information about the specific input data and weights used in the design of their recommender systems.
Platforms must publicly disclose the metrics they use to measure long-term user value.
Platforms must publicly disclose the metrics they use to evaluate product teams responsible for recommender system design.

Implementation discussion

Input data and weights

For each type of disclosure listed below, platforms should disclose the information as it applies across the entire user base, as well as with respect to individual user segments reflecting specific age, region, or other cohorts for which platforms specifically tailor their recommendations.

Input data: All the sources of raw information used in ranking should be disclosed. This could include item content and metadata, engagement history, user survey data, quality feedback from users, annotations from raters, user settings, profile and social graph data, context data (day, time, location, etc.) and other data sources.

In the case where input data to ranking is the output of another machine learning model, the input data to that model should also be included. An example of a (partial) disclosure of this type for Instagram is shown in Figure 4 below. Meta discloses some of the inputs to the machine learning models that feed into its rankings, but not all.

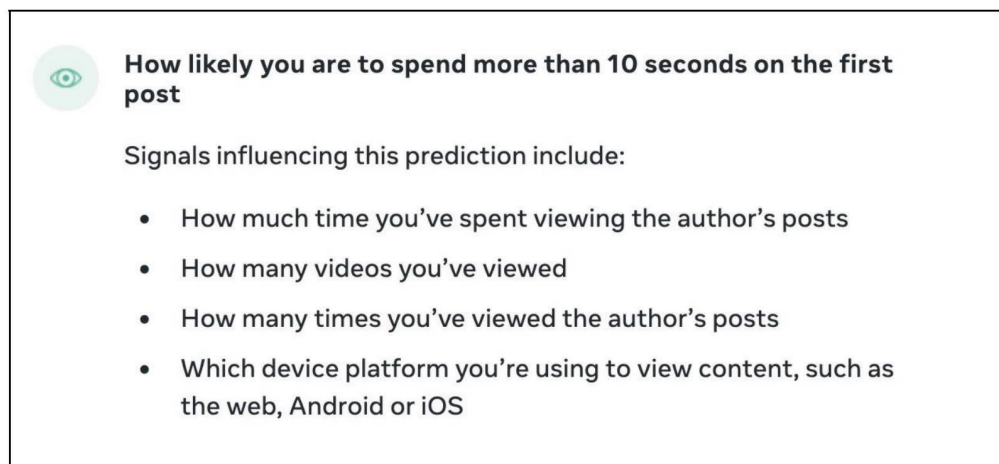


Figure 4. An image of Meta's disclosure about the Instagram Feed's recommender system. It shows some of the signals that factor into the recommender system's prediction of whether a user will spend more than 10 seconds on the feed's first post.¹⁴¹

The input data disclosure should also include data that comes from other products or features. For example, it should be disclosed whether or not YouTube uses user data from Google Search and whether the Facebook news feed uses ad engagement data for content selection. Knowing the universe of input data sources is necessary for independent experts to discern how recommender systems work.

Values and their weight quartiles: Most recommender systems rely on weights applied to some set of values in the system, and these weights reveal which values have greater or lesser impact on ranking. Platforms and services should report the complete list of values and their weights for the system as a whole (not for each individual user). Because weights are difficult to interpret numerically, and could be claimed by some parties to be trade secrets, the quartile of the weight should be reported instead of raw numeric weights. Achieving meaningful transparency of these weights is an intricate and technical challenge. Section VII(B) provides guidance to product teams about how they could most usefully disclose this information.

Metrics to measure long-term user value

The mandatory disclosure of certain categories of metrics used internally within platforms would incentivize platforms to develop and track these metrics, and it would enlighten cross-industry comparisons about how key metrics of interest are measured within different platforms. In particular, platforms should disclose the metrics they use to measure long-term user value. Platforms may approach the measurement of long-term value to users in a variety of different ways, including by tracking daily active use and specific forms of engagement over time, through user surveys of different kinds, and through other methods.

¹⁴¹ Meta, "Instagram Feed AI System." Under the guidance proposed in this section, all of the inputs into the recommender system would be disclosed instead of a subset.

This guideline does not require platforms to measure long-term value in a specific way, it merely requires public disclosure about how long-term user value is measured. Similar to the recommended holdout experiments discussed in Section V(C), the objective of requiring the disclosure of these metrics is to motivate consideration for users' long-term preferences and satisfaction in addition to platforms' and users' short-term gains. Public authorities can validate these metrics through independent cross-platform user surveys, which do not require any access to platform data. Regulators in the United Kingdom and Australia have already established such regular user experience monitoring.¹⁴²

For each metric that is published, the platform should disclose how the metric is calculated, which user segments it is calculated over, and the time window(s) over which it is calculated. User segments should include subgroups defined by demographic characteristics for which the platform collects data as well as those defined by quantitative thresholds of usage or engagement (for example, the top 10% or 1% of users as measured by specific forms of usage or engagement).

This disclosure requirement is purposefully limited to account for constraints created by the First Amendment in the US. Section VI(C) discusses additional metrics disclosures to be considered in other jurisdictions.

Metrics to evaluate product teams

Every company has its own way of setting objectives and measuring employee and team performance against those objectives. One method in common use within technology companies is known as “objectives and key results” or OKRs.¹⁴³ When using an OKR framework, each team or unit will establish an objective – a concrete and clearly defined goal – together with a small number measurable (and typically quantifiable) success criteria, the key results. Employee remuneration and recognition may be tied to achieving the KRs over some defined period, typically some number of quarters or a year. While not all companies use the OKR framework, most have some framework in place for focusing on a specific set of goals and a limited set of performance metrics that are used to measure progress towards those goals.

The set of key results or metrics used to measure product team performance for teams responsible for recommender systems reveal what platforms are most focused on achieving with their recommender systems at any given time. Platforms may track hundreds or thousands of different metrics that can be used to evaluate many different forms of engagement, revenue, and ad impressions, as well as quality and integrity metrics.¹⁴⁴ Requiring the disclosure of all of these metrics would provide a sea of information with no guide as to how the metrics are traded off against each other or which ones carry the most importance when platforms decide to make design changes.

¹⁴² eSafety Commissioner, “Australians’ Negative Online Experiences 2022”; Ofcom, “Experiences of Using Online Services.”

¹⁴³ See, e.g., Grove, *High Output Management*.

¹⁴⁴ Integrity Institute, “On Risk Assessment and Mitigation for Algorithmic Systems”; Public Interest Tech Lab, “Evaluating News Feed Ranking Experiments.”

Focusing on the metrics used to evaluate product teams provides a narrower window into what the platform views as most important. Making these metrics transparent should incentivize platforms to incorporate employee and team evaluation criteria that better align with user value. Product team metrics that are solely and consistently focused on engagement metrics and do not include metrics related to user value, satisfaction, or harm mitigation should be cause for alarm.

The metrics used to evaluate product teams can be viewed as sensitive from a business perspective because they may reveal information about product plans, roadmaps, or competitive strategies. For example, if a platform plans to launch a new feature during a particular quarter, a key result may be included related to the performance of that feature even before it has launched. The default should be that these metrics are publicly disclosed, but authorities implementing this requirement might consider incorporating an exception process in cases where they determine that heightened confidentiality is justified.

As with many disclosure requirements, a requirement to make product team metrics public would be strengthened if it also included provisions for auditing these metrics. It would be straightforward for platforms to maintain a true set of metrics used to judge team performance and report a different set of metrics publicly, since external parties have no other information to verify whether the metrics being reported are the true metrics. Ensuring that these metrics can be audited by an independent auditor would provide a check against platforms potentially gaming this requirement.

B. User Choices and Defaults

Many platforms have developed basic functionalities that allow users to shape recommendations. These functionalities include controls at the item level, system-wide restrictions on unwanted items, and options for how items are ranked.

Item-level controls typically permit users to hide or give feedback on an individual item, usually by tapping an adjacent button with a label like “hide”, “not interested,” or “show less like this.” System-wide restrictions include settings that limit the recommendation of potentially sensitive items, items with specific unwanted keywords, and items from unfamiliar languages. Finally, options for ranking let users toggle between a default feed and alternative feeds that rank items differently.

Some platforms have developed alternative ranking feeds, either voluntarily or to meet their compliance obligations under Article 38 of the EU DSA, which mandates that users be given an option for each of their recommender systems that is “not based on profiling.”¹⁴⁵ In practice, these feeds are chronological or non-personalized (i.e., items are ranked by recency or very general information about

¹⁴⁵ European Union, “Digital Services Act,” Article 38.

users).¹⁴⁶ Feeds like these, which are available to users but not set as the default, are often difficult to access and understand and therefore see little uptake.¹⁴⁷

Policies that mandate chronological or non-personalized feeds may be particularly counterproductive if users switch back to engagement-optimized feeds due to a poor user experience. This outcome may allow platforms to claim that users prefer engagement-optimized ranking, obscuring the spectrum of alternative designs, and it provides no incentive for platforms to improve their user experience beyond the baseline set by ranking for predicted engagement.

The user choice guidelines below focus instead on requiring alternative recommender systems that support long-term value to users and honor the preferences users set explicitly about items and types of items they do and do not want to see.

Guidelines:

Platforms must offer users an easily accessible choice of different recommender systems. At least one of these choices must be optimized to support long-term value to users.

Platforms must provide easily accessible ways for users to set their preferences about types of items to be recommended and to be blocked. Platforms must honor those preferences.

The first guideline would require platforms to offer at least one recommender system option that focuses on supporting long-term value to users, as discussed in the box above. Evidence from research and implementation has demonstrated a variety of approaches to designing recommender systems optimized for values other than predicted short-term engagement. This requirement does not mandate any specific approach, thereby giving platforms the freedom to design an alternative recommender system that supports long-term value to users in ways that reflect the unique characteristics of each platform. Yet, unlike a simple chronological or non-personalized feed, this option requires the platform to take deliberate steps to orient their systems around long-term value, whether by proactively soliciting long-term preferences from users, extrapolating from surveys or deliberative processes, relying on quality indicators selected by users, or through other means.

The second guideline operates at the level of individual items and categories of items. In furtherance of user agency, it requires platforms to abide by users' explicit indications about items they want to have recommended or blocked. A common complaint about existing user controls on social media

¹⁴⁶ Pershan and McCrosky, "No Perfect Solution to Platform Profiling Under Digital Services Act."

¹⁴⁷ Cunningham et al., "What We Know About Using Non-Engagement Signals," 14.

platforms is that they are difficult to access¹⁴⁸ and do not enhance users' sense of agency or control.¹⁴⁹ Even for the small fraction of users who engage with user controls, they often appear to have little or no effect on users' feeds.¹⁵⁰ If implemented effectively, these controls could be critical tools in helping users to exercise their long-term preferences – a user trying to reduce their consumption of unhealthy dieting videos, for example, could set preferences to limit this category of content from appearing in their feed. Requiring platforms to implement these controls in a robust and accessible manner is an important component in empowering users.

Implementation discussion

Choices and user controls are notoriously difficult to design effectively. Default settings will always have a much greater impact on overall user experience than settings users must choose themselves. Nevertheless, there is a vast body of work in human-computer interaction, user interface design, and behavioral economics from which platforms can draw to design their user choice and control architectures to be as effective and accessible as possible.¹⁵¹ For example, platform studies have shown how improving the accessibility and discoverability of controls can increase their uptake and positive reception among users.¹⁵²

Effective design will be crucial for the success of user controls. While the tools developed should be tailored to each platform, research has revealed some general design principles. Platforms should design controls knowing that users vary in what they want out of controls and the amount of knowledge they possess about how the platforms work.¹⁵³ Some users desire very granular controls (e.g., over individual pieces of content) while others would prefer coarser control over the inclusion of specific topics (e.g., weight loss content in their feeds).¹⁵⁴ Controls should be transparent and easily discoverable (or what the DSA calls “direct and easily accessible”).¹⁵⁵ Users should also be able to

¹⁴⁸ Platforms have also created features that attempt to explain why an individual item (e.g., an advertisement) was recommended to a user. Users can use this feature by tapping a button associated with an item that displays “Why am I seeing this post?” (or similar depending on the platform). However, these in-context explanations are not always helpful to users who desire transparency. See, e.g., Andreou et al., “Investigating Ad Transparency Mechanisms in Social Media”; Eslami et al., “Communicating Algorithmic Process in Online Behavioral Advertising”; Kim et al., “Why Am I Seeing This Ad?”

¹⁴⁹ Lukoff et al., “How the Design of YouTube Influences User Sense of Agency”; Zhang et al., “Monitoring Screen Time or Redesigning It?”

¹⁵⁰ Ibid.; Gak et al. “The Distressing Ads That Persist”; Ofcom, “Fewer than Half of Social Media Users Find Content Controls Effective.”

¹⁵¹ See, e.g., Calvo and Peters, *Positive Computing*; Davis et al., “Supporting Teens’ Intentional Social Media Use Through Interaction Design”; Desmet and Pohlmeier, “Positive Design”; Hassenzahl et al., “Needs, Affect, and Interactive Products”; Hassenzahl, *Experience Design*; Friedman and Hendry, *Value Sensitive Design*; Peters et al., “Designing for Motivation, Engagement and Wellbeing in Digital Experience.”

¹⁵² Cunningham et al., “What We Know About Using Non-Engagement Signals”, 14; Schnabel et al., “The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation.”

¹⁵³ Habib et al., “Identifying User Needs for Advertising Controls on Facebook”; Harper et al., “Putting Users in Control of Their Recommendations”; Jin et al., “Effects of Personal Characteristics in Control-Oriented User Interfaces”; Millecamp et al., “Controlling Spotify Recommendations”.

¹⁵⁴ Gak et al. “The Distressing Ads That Persist”; Habib et al., “Identifying User Needs for Advertising Controls on Facebook.”

¹⁵⁵ Schnabel et al., “The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation.”

understand how toggling options will affect their experience on a platform.¹⁵⁶ Platforms should aggregate available controls in an intuitive location that minimizes required navigation, such as by placing them in a prominent position within their settings.¹⁵⁷

Implementation details are crucial to the success of user controls. Platforms should present controls to users at key points in time when such choices feel most relevant and actionable. For example, platforms could incorporate preference selections during the signup process, at particular product use milestones,¹⁵⁸ or following significant updates that affect recommender systems. This approach would complement access through platform settings while ensuring broader awareness and adoption.

Applicability to minors

For minors in particular, there may be legitimate concerns that offering a choice of recommender systems does not go far enough to protect these users given the stage of their cognitive and social-emotional development. A better approach would be to require that minors receive an improved recommender system design by default.

Guideline:

By default, platforms must set minors' recommender systems to be optimized to support long-term value to these users. If platforms have insufficient information about long-term value to minors, they must default to non-personalized recommender systems.

The crux of this guideline still relies on the platform developing its own approach to optimizing for long-term user value, but sets that option as the default rather than as a choice. This should incentivize platforms to build an understanding of long-term value for users in this group by allowing users to explicitly indicate what they would like to see on the platform (while applying appropriate privacy protections to this data), by conducting user surveys or deliberative processes, by following guidance and evidence from public health authorities, or through other means. If the platform does not have a mechanism for identifying long-term value to minors, it must fall back to a non-personalized recommender system. Because of the demonstrated value of personalization to both users and platforms, establishing this as a backstop should provide an additional incentive for platforms to develop rigorous approaches to understanding long-term value to minors.

¹⁵⁶ For example, over 20% of ads on Facebook are run without topics specified by the advertiser. This means these ads will continue to be shown even if a user has requested not to be shown ads targeted using a particular topic. See Ali et al., “Problematic Advertising and Its Disparate Exposure on Facebook.”

¹⁵⁷ Habib et al., “Identifying User Needs for Advertising Controls on Facebook.”

¹⁵⁸ Redmiles et al., “Dancing Pigs or Externalities?”

C. Long-Term Holdout Experiments

Recommender systems are commonly optimized for short-term predicted engagement. Financial markets and investors typically prefer to see an ad-driven business grow its engagement and usage on a quarterly basis, which fuels a short-term focus.

Optimizing for longer term value, whether measured by user retention or satisfaction, is more difficult. Learning what makes users stick with the platform and increase their usage or satisfaction over time is a slow process; teams have to wait a year to find out from users what causes them to stick with a platform for that long. Running longer term experiments on a platform to find out what causes retention or satisfaction can be costly because it requires investment over time. The effects of platform changes that increase long-term retention or satisfaction can look like failures in the short run – daily active minutes might go down, but over months and years, user satisfaction might go up. This effect is depicted in Figure 5 below. Correlating short-term metrics with long-term metrics is difficult,¹⁵⁹ and it is risky for platforms to commit to a long-term experiment if the initial, short-term results suggest that the intervention is making things worse. Because of this, platforms tend to rely on short-term experiments to inform product design changes.

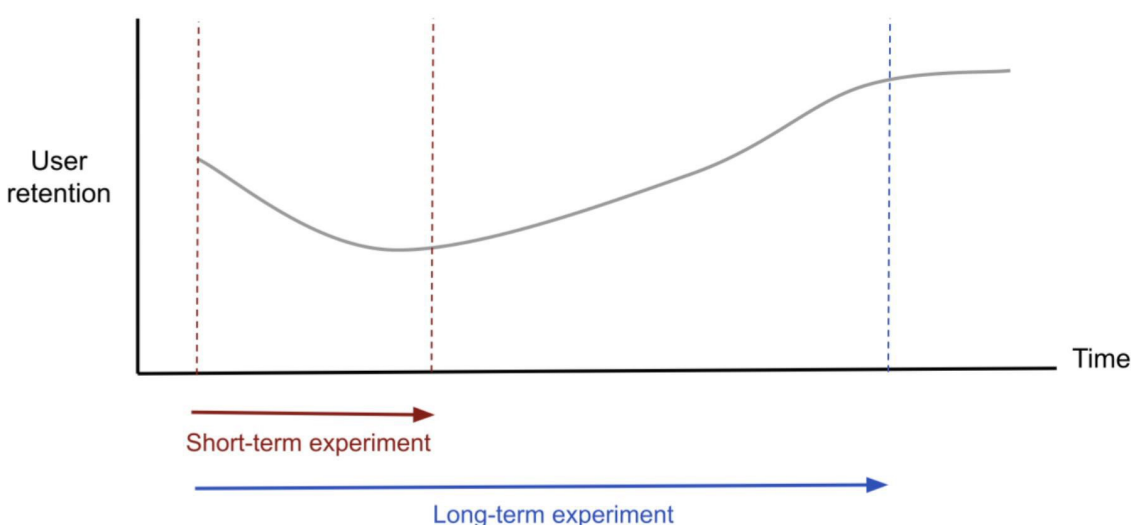


Figure 5. Example of a short-term versus long-term experimental effect. Platforms can lack sufficient incentives to conduct long-term experiments if early results indicate declines in user retention.

Online platforms may run thousands or tens of thousands of experiments each year to test out different design aspects of their systems, including changes to recommender system design.¹⁶⁰ Many of these experiments might last for days or weeks, after which time product teams evaluate their effectiveness against company-selected metrics and decide whether to maintain the changes, revert back, or continue experimenting. Over the course of a year, platforms might make dozens or hundreds of changes to the algorithms that power the recommendations each user sees.

¹⁵⁹ Cunningham et al., “What We Know About Using Non-Engagement Signals.”

¹⁶⁰ See, e.g., Donovan, “The Role of Experimentation at Booking.com.”

Given the frequency of experimentation, many platforms also maintain a holdout group – a group of users that are exempt from having design changes applied to their accounts, and who function as a control group for comparison with the rest of the user base.¹⁶¹ The holdout group size varies greatly from platform to platform. Most users never become aware that they are in a holdout group even though their user experience can vary significantly from all other users.

On some platforms, holdout experiments can be long-running, with users staying in the holdout group for years at a time. While most holdout experiments go undisclosed, in the past companies have voluntarily published the results of long-term holdouts demonstrating the effects of specific design choices on metrics of user welfare. For example, Meta and Twitter (now X) have previously shared the results of long-term holdouts that withheld users from receiving advertisements and personalized feeds, respectively.¹⁶² These holdouts were run continuously for years, underpinning their usefulness for understanding the design features they examined. Running holdouts of this length should be considered a best practice for platform design.

If long-term holdout experiments were to become more institutionalized and more available for public scrutiny, they could become powerful tools to shift platforms' incentives towards designs that optimize for long term user retention, value, and satisfaction. As explained in Section III(C), users often make choices to engage with content in the moment that are not indicative of what the users aspire to or prefer in the long run, or what makes them happy about their overall experience with a platform later on. When platforms optimize for short-term engagement, they exacerbate this dynamic. If platforms were required to demonstrate long-term user value, they would design their recommender systems differently. Requiring long-term holdout experiments and the public disclosure of their results will incentivize platforms to give more priority to long-term user retention, value, and satisfaction.

Guidelines:

Platforms must run long-term (12-month or longer) holdout experiments on a continuous basis.
Platforms must report the aggregate, anonymized results of the holdout experiments publicly.
Holdout experiments must be subject to an audit by an independent third party.

Focusing on long-term value to the user prioritizes long-run, forward-looking user preferences or aspirations (System 2) over short-run, impulsive user preferences (System 1). Platforms will be motivated to demonstrate positive outcomes when the holdout experiment results become public.

¹⁶¹ See, e.g., Pinterest Engineering, “How Holdout Groups Drive Sustainable Growth.”

¹⁶² Brynjolfsson et al., “The Consumer Welfare Effects of Online Ads”; Huszár et al., “Algorithmic Amplification of Politics on Twitter.”

They will want to show that users who received product updates throughout the year are more satisfied and more likely to stay on the platform compared to the holdout group that did not receive these changes. This creates a natural check against short-term thinking: if product changes boost immediate engagement but ultimately lead to users leaving, feeling dissatisfied, or experiencing harm, these problems will become clear when comparing the holdout group to the production group. Importantly, this incentive system works without requiring platforms to disclose their specific product changes, which they may want to keep private for competitive reasons. The key is that platforms will strive to prove that their collective changes over the year create better long-term value for users compared to the unchanged holdout experience.

Implementation discussion

Platforms must run long-term (12-month or longer) holdout experiments on a continuous basis and start a new holdout group each year. The holdout group's experience of the platform should be frozen around a snapshot of current platform design at the start of the experiment. The only exceptions should be for product changes whose objectives are specifically to reduce or prevent direct and immediate harms without increasing engagement or revenue (e.g., changes designed to reduce suicidal ideation or eating disorder exacerbation). Such changes may be applied in the same way to the holdout group and the production group.

The platform should establish a set of metrics that will be tracked for comparison between the holdout group and the production group. At a minimum, these should include:

- Retention metrics, for example, what percentage of users in each group are still daily active users (DAUs) at intervals over the course of the experiment (1 month, 6 months, 12 months)
- User surveys or interviews about their experiences (these can measure both positive and negative experiences as well as user satisfaction)
- Metrics that measure high-level constructs related to individual or systemic harms or benefits (well-being, conflict, etc.),¹⁶³ as defined by the platform

The results of all these measurements should be publicly reported in the aggregate (without revealing private data). See Section VII(A) for details about how holdout groups should be constructed.

For accountability, a third party should audit the holdout experiments on an annual basis. If the holdout infrastructure is neglected – e.g., if machine learning models that predict basic constructs like relevance are not re-trained in a timely fashion, as they would be in production – then the experience of the holdout group will be artificially poor compared to the production group. Platforms could also exploit their control over the holdout group to make the production recommender system seem as though it is drastically improving by comparison. For these reasons, the auditor needs to be able to assess the design of the recommender systems experienced by both the holdout group and the production group, as well as the composition of the holdout group compared to the overall user base.

¹⁶³ These should address common platform harms and use measures that have been validated by external researchers to the greatest extent possible. For a list of such measures, see Lubin et al., “Social Media Harm Abatement.”

VI. Global Policy Guidance

As described in Section IV, the US legal framework and jurisprudence related to corporate speech rights impose potential limitations on policies enacted legislatively in the US. Aspects of the EU DSA and the UK Online Safety Act, for example, would likely face legal challenges were they to be enacted in the US. This section offers additional policy guidance that may be more feasible to enact in jurisdictions outside the US, and that may be combined with the core policy guidance.

A. Public Content Transparency

Modern platforms are so large and complex that even the companies' own employees struggle to fully understand how their recommender systems affect users. Without being able to see and measure effects, it is very difficult to improve these systems and shift them away from chasing short-term engagement toward creating long-term value.¹⁶⁴ Meaningful improvement requires transparency about what content is actually being shown to users across the platform. However, most platforms currently provide very little detailed data about what content is most prevalent on their services, and in recent years multiple platforms have reduced the access they previously provided.¹⁶⁵

Guidelines:

Platforms must continuously publish a sample of the public content that is most highly disseminated on the platform and a sample of the public content that receives the highest engagement.

Platforms must continuously publish a representative sample of public content consumed during a typical user session on the platform at any given time.

There are two reasons for public content transparency: (1) to allow the public to validate companies' own reports about the prevalence of different kinds of content on their platforms; and (2) to raise awareness about potential harms and trends. Publishing samples of public content will not reveal content associated with some types of highly salient harms whose prevalence is too small to appear in the samples, but it will provide some indication of what content is disseminated based on a combination of algorithmic and other factors, providing an ongoing illustration of the output of

¹⁶⁴ Horwitz, *Broken Code*.

¹⁶⁵ Meta publishes a widely viewed content report for Facebook. It does not include many of the details recommended in this report, and most platforms do not publish anything similar. See Meta, "Widely Viewed Content Report." Moreover, many major platforms have recently deprecated or restricted usage of public channels for accessing data, including Meta's retirement of CrowdTangle, X's price increase for access to its application programming interface, and Reddit's changes to its terms for similar access. See Hickey et al., "Public Data Access Programs"; Gotfredsen and Dowling, "Meta Is Getting Rid of CrowdTangle — and Its Replacement Isn't As Transparent or Accessible"; Perez, "Reddit Locks down Its Public Data in New Content Policy, Says Use Now Requires a Contract"; Stokel-Walker, "Twitter's \$42,000-per-Month API Prices Out Nearly Everyone."

recommender systems. For privacy reasons, these samples should be limited to public content when broadly disseminated, although policymakers could require that access to non-public content be granted if it is limited to appropriate regulators or authorized reviewers.¹⁶⁶

Implementation discussion

If adopted, these guidelines would entail a significant expansion of transparency by social media platforms. While it is already standard practice for these companies to publish quarterly transparency reports about, for example, the enforcement of their content guidelines and government requests for information,¹⁶⁷ these documents provide little information about the top publicly available content. One exception is Meta’s “Most Widely Viewed Content Report,”¹⁶⁸ which discloses the top 20 domains, links, pages, and posts on Facebook over each three month period. Figure 6 shows one example entry. This information is very sparse and therefore of limited utility to external observers.

Rank	Post Link	Post Image	Content Viewers
1	facebook.com/880612957258432		57.1M

Figure 6. Example of a top content disclosure from Meta’s “Most Widely Viewed Content Report.” According to the report, this post received the most views on Facebook in the third quarter of 2024.

The transparency efforts envisioned by the two guidelines would result in sharing much more detailed and complete data about the top public content. The first guideline requires platforms to continuously

¹⁶⁶ At present, there is no settled definition of “publicly accessible platform data.” The DSA includes provisions to enable independent research with publicly accessible platform data, but has not established a specific definition. See European Union, “Digital Services Act,” Article 40. CrowdTangle included numerical thresholds to define public availability including data made by a public page, group, or (possibly) verified public person, who had more than 110,000 likes (or was added to a CrowdTangle list). See Garmur et al., “CrowdTangle Platform and API.” Work is underway to establish common definitions for publicly accessible platform data. See Knight-Georgetown Institute, “The Gold Standard for Publicly Available Platform Data.”

¹⁶⁷ See Google, “Google Transparency Report”; Meta, “Transparency Reports”; Snap, “Snapchat Transparency Report”; TikTok, “Transparency Center”; X, “X Transparency Center.”

¹⁶⁸ Meta, “Widely Viewed Content Report.”

publish a sample of the most widely-disseminated content on the platform (i.e., that reaches the widest audience). Thresholds for what fraction of content the sample draws from (e.g., top 1% or 5%) could usefully be standardized and specified in regulation.

This guideline also requires platforms to maintain a dynamically-updated list of the most-engaged-with items on the platform at any given time, defined over a relatively short time period (such as the last seven days). The definition of most-engaged-with will require the platform to identify the forms of engagement (clicks, likes, reshares, etc.) that are most relevant for distribution on their specific platform and publish content samples and associated engagement data for each individual form. For example, watch time may be the most meaningful form of engagement on a video streaming service and essentially irrelevant on a platform that does not support video. For such a service, the platform would be expected to publish content samples based on the highest watch time. If platforms are required to disclose the weight quartiles of their recommender system inputs as proposed above in Section V(A), the forms of engagement reflected in the inputs in the upper quartile would be an appropriate set for selecting and publishing most-engaged-with content samples.

The second guideline requires platforms to publish a dynamically-updated sample of public content that a user might see during a session of typical length on the platform. This sample would provide an indication of the representative user experience and content mix during a session – how much content from followed accounts or subscriptions, how much of different media types (images, video, etc.), or how much from different content categories, for example. This could be published as a list of recommended items that a user may see during an average-length session, along with metadata describing each item (such as aggregate engagement, date and type of publisher, type of media, etc.).

B. User Defaults

Section V(B) recommends that platforms provide users with a choice of different recommender systems, where at least one choice is optimized to support long-term value to users. It also recommends that, at minimum, minors have the design that supports long-term value set as the default.

In jurisdictions where the legal framework allows it, the default requirement can be expanded to apply to the entire user population, rather than limited to minors. Defaulting all users into designs optimized for long-term value – even if it means sacrificing short-term engagement – should be understood as a best practice.

Guideline:

By default, platforms must optimize users' recommender systems to support long-term user value.

As with the specific guideline for minors in Section V(B) above, optimizing for long-term value for the whole user population would spur platforms to build the processes needed to gather and understand what creates user value and satisfaction.

Adopting this requirement would fully unlock the combined potential of the three sets of guidelines outlined in this report: transparency, long-term holdout experiments, and user choices and defaults. If default feeds are optimized for long-term user value, then long-term holdout experiments should demonstrate that the production group is deriving more satisfaction than the holdout group. Similarly, the published metrics that measure long-term value and individual harm should begin to reflect the steps taken to optimize default feeds across the platform.

C. Metrics and Measurement

As outlined in Section III(A), there are a number of different vectors of harm in which recommender systems may play a role depending on the platform: unwanted or harmful content, product usage, contact, or usage of personal information. Platforms that operate at a scale to run frequent A/B tests for engagement should track metrics associated with the potential harms to at-risk populations that are relevant to their platforms. Surveys and other long-term value measures such as the ones outlined earlier in this document would be suitable for such measurement.

Guideline:

Platforms must measure the aggregate harms to at-risk populations that result from recommender systems and publicly disclose the results of those measurements.

The key aspect of these measurements is that they are designed to evaluate effects on populations and not on individuals – reflective of the types of harms that system architecture is capable of causing.¹⁶⁹ Because these types of assessments and metrics operate at the population level, they can generally be revealed without implicating the privacy of individual users.

The specific measurements needed will vary from platform to platform. Examples include:

- *Unwanted or harmful contact and content.* These can be measured through surveys about negative experiences,¹⁷⁰ by tracking user behaviors that indicate negative experiences (e.g.,

¹⁶⁹ Lubin et al., “Social Media Harm Abatement”; Lubin and Gilbert, “Accountability Infrastructure.”

¹⁷⁰ For example, Meta has conducted internal research using its “Bad Experiences and Encounters Framework (BEEF)” survey to ask minors about various online harms, including hate speech and unwanted sexual advances. The findings of this research were not made public until they were disclosed as part of ongoing litigation with state Attorneys General. In the absence of transparency about internal survey results like this one, outside groups have launched projects such as the Neely Social Media Index to ask users about their negative experiences online. See Horwitz, “His Job Was to Make Instagram Safe for Teens.”; Fast et al., “Unveiling the Neely Ethics & Technology Indices.”

hiding content, blocking users, reporting content), or by measuring engagement with content that violates platform policies or that is predicted to violate platform policies.

- *Unwanted or harmful usage.* To track harmful usage or sleep effects, platforms might track the percentage of users who use the platform for an excessive number of hours per day, during school and nighttime hours, or very frequently. Platforms can combine those metrics with survey data about unwanted usage, sleep, regret, and activity displacement to understand when users themselves feel that their usage is problematic.
- *Systemic harms.* These vary widely. For example, measurements of conflict or polarization on a platform might combine user surveys with engagement metrics that measure engagement with content identified as toxic towards outgroups.¹⁷¹

Platform-wide measurements might provide additional insights across multiple of these harm categories. For example, platforms might survey older users to understand the experiences they had on the platform as younger users.

VII. Best Practices for Product Teams

This section provides technical guidance about how product teams can implement several of the guidelines described above.

A. How to Construct Holdout Groups

When conducting holdout group experiments, the size of the holdout group should be sufficiently large such that any significant movements in the key metrics can be measured with confidence.¹⁷² This will depend both on the overall number of users on the platform and the magnitude of typical annual changes. For example, platforms with more users will be able to allocate a smaller percentage of the user base to the holdout and still achieve statistically significant results. And if the typical movement of the metrics, over the course of the year, is a few percentage points, then the holdout should be large enough to detect effect sizes of a percentage point with high confidence. Platforms should publicly report the absolute and relative size of the holdout group compared to the overall user base.

The composition of the holdout group is critical to the effectiveness of the experiment. At a minimum, the holdout group should be representative of the overall user base across key demographics for which the platform collects data. For example, if the platform directly collects age and geographic region information from its users, the holdout group should be representative of the overall user base across those demographics. The composition of the holdout group should also be representative of

¹⁷¹ Stray, “Dependent Variables.”

¹⁷² Allen and Lawson, “Proposal for an Assessment of Risk Mitigations for Algorithmic Amplification of Disinformation.”

the engagement levels of the overall user base, i.e., the distribution of engagement levels across the user base should be reflected in the holdout group.

To provide more comprehensive insights, platforms should oversample user subgroups for which there is evidence of disproportionate risk of harm on the platform (e.g., minors). At minimum, this should include cohorts exhibiting the most significant engagement (e.g., the top 1% or top 0.1% of highest usage). But depending on the product, other subgroups are likely to be appropriate for oversamples, such as minors or other at-risk demographics for which segmentation is possible. Platforms should publicly report how they assemble a representative sample in the holdout group, which subgroups they oversample, and why.

B. How to Disclose Recommender System Weights

All recommender systems are ultimately driven by some set of human-chosen weights, i.e., numeric settings on high-level parameters. In early systems, such weights (or coefficients) were directly applied to content features. For example, the timeliness of an item or the number of comments on an item would be factored directly into the ranking of that item. Today, weights are more commonly applied to the output of multiple machine learning models. For example, weights might be applied to the predicted probability of user engagements of various types, or the probability of various content attributes such as clickbait, nudity, etc. Other systems use hand-built rules to label the value of past engagement for each user, and then train models to generalize these patterns so as to predict the value of showing a particular item to a particular user at a particular time. In all cases there is some set of weights that ultimately drives training of machine learning models.

Whatever type of values or signals or inputs are ultimately weighted, platforms and services should report the complete list of values and their weights. These should be disclosed for the system as a whole, not each individual user. Because weights are difficult to interpret numerically, and could be claimed by some parties to be trade secrets, the quartile of the weight on each value as well as the distribution of the weights should be reported instead of raw numeric values.

Depending on the design of the system, the weight quartiles may be difficult to interpret. To take a simple example, suppose predicted time spent on a piece of content is a value in the ranking system. In that case, a platform could decrease the weight reported while maintaining the same ranking score by converting its time measurements from minutes to seconds. This would artificially reduce the weight assigned to predicted time spent to 1/60th of the weight in minutes despite having no change to the ranking. Similar issues could arise with other values.

To facilitate interpretation and mitigate the possibility of platforms gaming these disclosures by scaling values up or down before reporting them, for platforms that use a linear value model, the weights should be normalized by the standard deviation of the signal. To take another simple example, suppose the value function is $(\text{weight}_1 * \text{signal}_1) + (\text{weight}_2 * \text{signal}_2)$. In this case, the weights that are reported must be $(\text{weight}_1 * \text{standard deviation of signal}_1)$ and $(\text{weight}_2 * \text{standard deviation of signal}_2)$.

The application of these weights may vary depending on the user segment or other factors (for example, less weight may be applied to certain values for new accounts versus long-time users). Such factors should be reported as well.

As discussed in Sections II and III, recommender systems may be designed using a variety of different foundational technical approaches. Some assemble a value model using a linear combination of terms, whereas others measure value based on machine learning outputs. The disclosure of weights recommended in Section V(A) would take different forms depending on the underlying technical approach.

A platform that uses a linear value model would report the value model, normalized as described above. If the platform uses different linear value models based on some conditional logic about the user, the item, or some other factors (e.g., a decision tree), the platform would report the different value models normalized as above, along with the conditional logic the platform uses to determine which value model is used for a given user or item.

If the primary measure of value by which a platform ranks content is the output of a machine learning model (e.g., a neural network), the platform would report the way that it operationalizes value when training the model. For example:

- If the model was trained to predict value using supervised learning, the platform would report how it constructed the labels on which the model was trained.
- If the model is trained using reinforcement learning, the platform would report how it constructed the reward function.

VIII. Conclusion

Recommender systems play an integral role in shaping online experiences, yet their design and implementation often prioritize short-term engagement over long-term user satisfaction and societal well-being. Better recommender systems are possible. Research has demonstrated that the landscape of alternative design approaches is significant. As proposals to regulate recommender system design mature, there is an opportunity to incentivize innovative designs that prioritize long-term user value and high-quality user experiences.

This report has established key policy and design guidelines to help make this a reality. By adopting detailed transparency measures, user choices and strong defaults, and accountability through long-term experiments, policymakers and product designers can help to foster vibrant online spaces while mitigating harms.

Bibliography

- 5Rights Foundation. “Pathways: How Digital Design Puts Children at Risk.” 5Rights Foundation, September 17, 2021. <https://5rightsfoundation.com/resource/pathways-how-digital-design-puts-children-at-risk/>.
- Agan, Amanda Y., Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. “Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias.” Working Paper Series. National Bureau of Economic Research, February 2023. <https://doi.org/10.3386/w30981>.
- Agarwal, Arpit, Nicolas Usunier, Alessandro Lazaric, and Maximilian Nickel. “System-2 Recommenders: Disentangling Utility and Engagement in Recommendation Systems via Temporal Point-Processes.” In *FAccT ’24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2024. <https://doi.org/10.1145/3630106.3659004>.
- Ali, Muhammad, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, and Piotr Sapiezynski. “Problematic Advertising and Its Disparate Exposure on Facebook.” In *Proceedings of the 32nd USENIX Conference on Security Symposium*, 5665–82. SEC ’23, 2023.
- Allcott, Hunt, Matthew Gentzkow, and Lena Song. “Digital Addiction.” *American Economic Review* 112, no. 7 (July 2022): 2424–63. <https://doi.org/10.1257/aer.20210867>.
- Allen, Jeff, and Abigail Lawson. “Proposal for an Assessment of Risk Mitigations for Algorithmic Amplification of Disinformation, the Role of Platform Business Models & Demonetization.” European Digital Media Observatory, June 6, 2024. <https://edmo.eu/edmo-news/report-proposal-for-an-assessment-of-risk-mitigations-for-algorithmic-amplification-of-disinformation-the-role-of-platform-business-models-demonetization/>.
- Alonzo, Rea, Junayd Hussain, Saverio Stranges, and Kelly K. Anderson. “Interplay between Social Media Use, Sleep Quality, and Mental Health in Youth: A Systematic Review.” *Sleep Medicine Reviews* 56 (April 1, 2021): 101414. <https://doi.org/10.1016/j.smrv.2020.101414>.
- Amnesty International. “Driven into Darkness: How TikTok’s ‘For You’ Feed Encourages Self-Harm and Suicidal Ideation.” Amnesty International, November 7, 2023. <https://www.amnesty.org/en/documents/pol40/7350/2023/en/>.
- Andreou, Athanasios, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. “Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook’s Explanations.” In *Networking and Distributed Systems Security*, 2018. <https://hal.science/hal-01955309/>.
- Austin, Mackenzie, and Max Levy. “Speech Certainty: Algorithmic Speech and the Limits of the First Amendment.” *Stanford Law Review* 77, no. 1 (January 10, 2025).

<https://www.stanfordlawreview.org/print/article/speech-certainty-algorithmic-speech-and-the-limits-of-the-first-amendment/>.

Balkin, Jack M. “Free Speech Versus the First Amendment.” *UCLA Law Review* 70, no. 1206 (2023).
<https://www.uclalawreview.org/free-speech-versus-the-first-amendment/>.

Bandy, Jack, and Tomo Lazovich. “Exposure to Marginally Abusive Content on Twitter.” *Proceedings of the International AAAI Conference on Web and Social Media* 17, no. 1 (June 2, 2023): 24–33.
<https://doi.org/10.1609/icwsm.v17i1.22123>.

Bavel, Jay J. Van, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. “How Social Media Shapes Polarization.” *Trends in Cognitive Sciences* 25, no. 11 (November 1, 2021): 913–16.
<https://doi.org/10.1016/j.tics.2021.07.013>.

Bengani, Priyanjana. “What’s Right and What’s Wrong with Optimizing for Engagement.” *Understanding Recommenders*, April 27, 2022.
<https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851>.

Besbes, Omar, Yash Kanoria, and Akshit Kumar. “The Fault in Our Recommendations: On the Perils of Optimizing the Measurable.” May 7, 2024. <https://doi.org/10.48550/arXiv.2405.03948>.

Brailovskaia, Julia, Verena J. Swarlik, Georg A. Grethe, Holger Schillack, and Jürgen Margraf. “Experimental Longitudinal Evidence for Causal Role of Social Media Use and Physical Activity in COVID-19 Burden and Mental Health.” *Zeitschrift Fur Gesundheitswissenschaften = Journal of Public Health*, September 2, 2022, 1–14. <https://doi.org/10.1007/s10389-022-01751-x>.

Brautsch, Louise AS., Lisbeth Lund, Martin M. Andersen, Poul J. Jennum, Anna P. Folker, and Susan Andersen. “Digital Media Use and Sleep in Late Adolescence and Young Adulthood: A Systematic Review.” *Sleep Medicine Reviews* 68 (April 1, 2023). <https://doi.org/10.1016/j.smrv.2022.101742>.

Bryant, Miranda. “Instagram Actively Helping Spread of Self-Harm among Teenagers, Study Finds.” *The Guardian*, November 30, 2024.
<https://www.theguardian.com/technology/2024/nov/30/instagram-actively-helping-to-spread-of-self-harm-among-teenagers-study-suggests>.

Brynjolfsson, Erik, Avinash Collis, Asad Liaqat, et al. “The Consumer Welfare Effects of Online Ads: Evidence from a 9-Year Experiment.” National Bureau of Economic Research, August 2024.
<https://doi.org/10.3386/w32846>.

Budzinski, Oliver, Sophia Gänßle, and Nadine Lindstädt-Dreusicke. “Data (r)Evolution - The Economics of Algorithmic Search and Recommender Services.” *Ilmenau Economics Discussion Papers*, 2021.
<https://ideas.repec.org/p/zbw/tuiedp/148.html>.

California. “AB 587.” 2022.
https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202120220AB587.

- — —. “California Age-Appropriate Design Code Act.” 2022.
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB2273.
- — —. “Protecting Our Kids from Social Media Addiction Act.” 2024.
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB976.
- Calvo, Rafael A., and Dorian Peters. *Positive Computing: Technology for Wellbeing and Human Potential*. The MIT Press, 2014. <https://doi.org/10.7551/mitpress/9764.001.0001>.
- Carter, Ben, Philippa Rees, Lauren Hale, Darsharna Bhattacharjee, and Mandar S. Paradkar.
“Association Between Portable Screen-Based Media Device Access or Use and Sleep Outcomes: A Systematic Review and Meta-Analysis.” *JAMA Pediatrics* 170, no. 12 (December 1, 2016): 1202–8. <https://doi.org/10.1001/jamapediatrics.2016.2341>.
- Casey, B.J., Rebecca M. Jones, and Todd A. Hare. “The Adolescent Brain.” *Annals of the New York Academy of Sciences* 1124, no. 1 (2008): 111–26. <https://doi.org/10.1196/annals.1440.010>.
- Center for Countering Digital Hate. “Deadly by Design: TikTok Pushes Harmful Content Promoting Eating Disorders and Self-Harm into Young Users’ Feeds.” Center for Countering Digital Hate, December 15, 2022. <https://counterhate.com/research/deadly-by-design/>.
- Chen, Yixin, Yue Fu, Zeya Chen, Jenny Radesky, and Alexis Hiniker. “The Engagement-Prolonging Designs Teens Encounter on Very Large Online Platforms.” January 29, 2025.
<https://doi.org/10.48550/arXiv.2411.12083>.
- Cho, Hyunsung, DaEun Choi, Donghwi Kim, Wan Ju Kang, Eun Kyoung Choe, and Sung-Ju Lee.
“Reflect, Not Regret: Understanding Regretful Smartphone Use with App Feature-Level Analysis.” *Proc. ACM Human-Computer Interaction*. (October 18, 2021): 456:1-456:36.
<https://doi.org/10.1145/3479600>.
- Christakopoulou, Konstantina, Madeleine Traverse, Trevor Potter, et al. “Deconfounding User Satisfaction Estimation from Response Rate Bias.” In *Proceedings of the 14th ACM Conference on Recommender Systems*, 450–55. RecSys ’20, 2020.
<https://doi.org/10.1145/3383313.3412208>.
- Christakopoulou, Konstantina, Can Xu, Sai Zhang, et al. “Reward Shaping for User Satisfaction in a REINFORCE Recommender.” In *Reinforcement Learning for Real Life (RL4RealLife) Workshop*, 2021. <https://doi.org/10.48550/arXiv.2209.15166>.
- Colorado. “SB158,” 2024.
https://leg.colorado.gov/sites/default/files/documents/2024A/bills/2024a_158_01.pdf.
- Common Sense Media. “Constant Companion: A Week in the Life of a Young Person’s Smartphone Use,” September 26, 2023.
<https://www.commonsensemedia.org/research/constant-companion-a-week-in-the-life-of-a-young-persons-smartphone-use>.

- — —. “Who Is the ‘You’ in YouTube?: Missed Opportunities in Race and Representation in Children’s YouTube Videos.” Common Sense Media, June 14, 2022.
<https://www.commonsensemedia.org/research/who-is-the-you-in-youtube-missed-opportunities-in-race-and-representation-in-childrens-youtube-videos>.
- Costello, Nancy, Rebecca Sutton, Madeline Jones, et al. “Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth from the Dangers of Social Media.” *American Journal of Law & Medicine* 49, no. 2–3 (July 2023): 135–72. <https://doi.org/10.1017/amj.2023.25>.
- Crone, Eveline A., and Elly A. Konijn. “Media Use and Brain Development during Adolescence.” *Nature Communications* 9, no. 1 (February 21, 2018): 1–10. <https://doi.org/10.1038/s41467-018-03126-x>.
- Cunningham, Tom, Sana Pandey, Leif Sigerson, et al. “What We Know About Using Non-Engagement Signals in Content Ranking.” arXiv, February 9, 2024. <https://doi.org/10.48550/arXiv.2402.06831>.
- Davis, Katie, Petr Slovak, Rotem Landesman, et al. “Supporting Teens’ Intentional Social Media Use Through Interaction Design.” In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 2023.
<https://dl-acm-org.proxy.library.georgetown.edu/doi/10.1145/3585088.3589387>.
- Desmet, Pieter M. A., and Anna E Pohlmeier. “Positive Design: An Introduction to Design for Subjective Well-Being.” *International Journal of Design* 7, no. 3 (2013).
<http://www.ijdesign.org/index.php/IJDesign/article/view/1666>.
- Donnelly, Robert, Ayush Kanodia, and Ilya Morozov. “Welfare Effects of Personalized Rankings.” *Marketing Science* 43, no. 1 (2024): 92–113.
- Donovan, Nicola. “The Role of Experimentation at Booking.Com.” *Click. Magazine*, September 2, 2019.
<https://partner.booking.com/en-us/click-magazine/industry-perspectives/role-experimentation-bookingcom>.
- eSafety Commissioner. “Australians’ Negative Online Experiences 2022.” Australia eSafety Commissioner, 2022.
<https://www.esafety.gov.au/research/australians-negative-online-experiences-2022>.
- Eslami, Motahhare, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. “Communicating Algorithmic Process in Online Behavioral Advertising.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. CHI ’18, 2018.
<https://doi.org/10.1145/3173574.3174006>.
- European Commission. “Supervision of the Designated Very Large Online Platforms and Search Engines under DSA.” Accessed February 28, 2025.
<https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>.

- European Union. “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act),” 2022. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>.
- Evans, J.S.B.T., and Keith Frankish. *In Two Minds: Dual Processes and Beyond*. Oxford University Press, 2009. <https://doi.org/10.1093/acprof:oso/9780199230167.001.0001>.
- Fast, Nathanael, Juliana Schroeder, Ravi Iyer, and Matt Motyl. “Unveiling the Neely Ethics & Technology Indices.” *Designing Tomorrow*, June 22, 2023. <https://psychoftech.substack.com/p/unveiling-the-neely-ethics-and-technology>.
- Feiner, Lauren. “Meta Sued by 42 Attorneys General Alleging Facebook, Instagram Features Are Addictive and Target Kids.” *CNBC*, October 24, 2023. <https://www.cnn.com/2023/10/24/bipartisan-group-of-ags-sue-meta-for-addictive-features.html>
- Florida. “SB 7072,” 2021. <https://www.flsenate.gov/Session/Bill/2021/7072>.
- Friedman, Batya, and David G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019. <https://mitpress.mit.edu/9780262039536/value-sensitive-design/>.
- Gak, Liza, Seyi Olojo, and Niloufar Salehi. “The Distressing Ads That Persist: Uncovering The Harms of Targeted Weight-Loss Ads Among Users with Histories of Disordered Eating.” In *Proceedings of the ACM Human-Computer Interaction*, 2022. <https://doi.org/10.1145/3555102>.
- Garmur, Matt, Gary King, Zagreb Mukerjee, Nate Persily, and Brandon Silverman. “CrowdTangle Platform and API.” *Harvard Dataverse*, 2019. <https://doi.org/10.7910/DVN/SCCQYD>.
- Glotfelter, Angela. “Algorithmic Circulation: How Content Creators Navigate the Effects of Algorithms on Their Work.” *Computers and Composition* 54 (December 1, 2019): 102521. <https://doi.org/10.1016/j.compcom.2019.102521>.
- Goodrow, Cristos. “On YouTube’s Recommendation System.” *Inside YouTube*, September 15, 2021. <https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/>.
- Google. “Activity Controls.” Google Account. Accessed January 28, 2025. <https://myactivity.google.com/activitycontrols>.
- — —. “Google Transparency Report,” 2024. <https://transparencyreport.google.com/>.
- — —. “Manage Your Recommendations and Search Results.” YouTube Help Center. Accessed January 28, 2025. <https://web.archive.org/web/20230312005351/https://support.google.com/youtube/answer/6342839>.
- Gotfredsen, Sarah Grevy, and Kaitlyn Dowling. “Meta Is Getting Rid of CrowdTangle — and Its Replacement Isn’t As Transparent or Accessible.” *Proof*, July 9, 2024. <https://www.proofnews.org/meta-is-getting-rid-of-crowdtangle-and-its-replacement-isnt-as-transparent-or-accessible/>.
- Grove, Andrew S. *High Output Management*. Vintage, 1995.

- Guess, Andrew M., Neil Malhotra, Jennifer Pan, et al. “How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?” *Science* 381, no. 6656 (July 28, 2023): 398–404. <https://doi.org/10.1126/science.abp9364>.
- Habib, Hana, Sarah Pearman, Ellie Young, Ishika Saxena, Robert Zhang, and Lorrie Falth Cranor. “Identifying User Needs for Advertising Controls on Facebook.” In *Proceedings of the ACM Human-Computer Interaction Conference*, 2022. <https://doi.org/10.1145/3512906>.
- Harper, F. Maxwell, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. “Putting Users in Control of Their Recommendations.” In *Proceedings of the ACM Conference on Recommender Systems*, 3–10. RecSys ’15, 2015. <https://doi.org/10.1145/2792838.2800179>.
- Hassenzahl, Marc. *Experience Design: Technology for All the Right Reasons*. Springer International Publishing, 2010. <https://doi.org/10.1007/978-3-031-02191-6>.
- Hassenzahl, Marc, Sarah Diefenbach, and Anja Goritz. “Needs, Affect, and Interactive Products – Facets of User Experience.” *Interacting with Computers* 2, no. 5 (September 2010). <https://academic-oup-com.proxy.library.georgetown.edu/iwc/article/22/5/353/684432>.
- Haugen, Frances. “Providing Negative Feedback Should Be Easy (And Why This Would Be Game Changing for Integrity),” 2019. https://www.documentcloud.org/documents/21602498-tier1_rank_pr_0919/.
- Hawaii. “Anti-Big-Tech Censorship Act,” 2023. https://www.capitol.hawaii.gov/session/archives/measure_indiv_Archives.aspx?billtype=HB&billnumber=1333&year=2024.
- Hickey, Cameron, Kaitlyn Dowling, Isabella Navia, and Claire Pershan. “Public Data Access Programs - A First Look.” Mozilla Foundation, August 8, 2024. <https://foundation.mozilla.org/en/blog/new-research-tech-platforms-data-access-initiatives-vary-widely/>.
- Hilbert, Martin, Drew P. Cingel, Jingwen Zhang, et al. “#BigTech @Minors: Social Media Algorithms Quickly Personalize Minors’ Content, Lacking Equally Quick Protection,” December 19, 2023. <https://doi.org/10.2139/ssrn.4674573>.
- Hödl, Tatjana, and Thomas Myrach. “Content Creators Between Platform Control and User Autonomy.” *Business & Information Systems Engineering* 65, no. 5 (October 1, 2023): 497–519. <https://doi.org/10.1007/s12599-023-00808-9>.
- Horwitz, Jeff. *Broken Code: Inside Facebook and the Fight to Expose Its Harmful Secrets*. Doubleday, 2023.
- — —. “His Job Was to Make Instagram Safe for Teens. His 14-Year-Old Showed Him What the App Was Really Like.” *Wall Street Journal*, November 3, 2023. <https://www.wsj.com/tech/instagram-facebook-teens-harassment-safety-5d991be1>.

- Hosseinmardi, Homa, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J. Watts. “Causally Estimating the Effect of YouTube’s Recommender System Using Counterfactual Bots.” *Proceedings of the National Academy of Sciences* 121, no. 8 (February 20, 2024): 1–8. <https://doi.org/10.1073/pnas.2313377121>.
- Huang, Saffron, Divya Siddarth, Liane Lovitt, et al. “Collective Constitutional AI: Aligning a Language Model with Public Input.” In *FACCT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417. Association for Computing Machinery, 2024. <https://doi.org/10.1145/3630106.3658979>.
- Husovec, Martin. “The DSA Newsletter #6.” *Tech Notes*, September 26, 2024. <https://husovec.eu/2024/09/the-dsa-newsletter-6/>.
- Huszár, Ferenc, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. “Algorithmic Amplification of Politics on Twitter.” *Proceedings of the National Academy of Sciences* 119 (January 4, 2022). <https://doi.org/10.1073/pnas.2025334119>.
- Integrity Institute. “On Risk Assessment and Mitigation for Algorithmic Systems.” Integrity Institute, February 29, 2024. <https://integrityinstitute.org/news/institute-news/risk-assessment>.
- Iyer, Ravi, Matt Motyl, and Nathanael Fast. “How User Experience Metrics Complement ‘Content That Requires Enforcement.’” *Designing Tomorrow*, July 19, 2023. <https://psychoftech.substack.com/p/how-user-experience-metrics-complement>.
- Jigsaw. “Perspective API - How It Works.” Accessed January 7, 2025. <https://perspectiveapi.com/how-it-works/>.
- Jin, Yucheng, Bruno Cardoso, and K. Verbert. “How Do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations?” In *IntRS@RecSys 1*, 2017. <https://www.semanticscholar.org/paper/How-Do-Different-Levels-of-User-Control-Affect-Load-Jin-Cardoso/5a5eab5cf7b5b18f88b23a65f50b58de2c23e260>.
- Kim, Tami, Kate Barasz, and Leslie K John. “Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness.” *Journal of Consumer Research* 45, no. 5 (February 1, 2019): 906–32. <https://doi.org/10.1093/jcr/ucy039>.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. “The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization.” *Management Science* 70, no. 9 (November 2023): 6336–55. <https://doi.org/10.1287/mnsc.2022.03683>.
- Knight-Georgetown Institute. “The Gold Standard for Publicly Available Platform Data.” Accessed February 28, 2025. <https://kgi.georgetown.edu/expert-working-groups/gold-standard-expert-working-group/gold-standard-faq/>.

- Kubin, Emily, and Christian von Sikorski. “The Role of (Social) Media in Political Polarization: A Systematic Review.” *Annals of the International Communication Association* 45, no. 3 (July 3, 2021): 188–206. <https://doi.org/10.1080/23808985.2021.1976070>.
- LeBourgeois, Monique K., Lauren Hale, Anne-Marie Chang, Lameese D. Akacem, Hawley E. Montgomery-Downs, and Orfeu M. Buxton. “Digital Media and Sleep in Childhood and Adolescence.” *Pediatrics* (November 1, 2017): 92–96. <https://doi.org/10.1542/peds.2016-1758J>.
- LinkedIn. “Use LinkedIn Reactions.” LinkedIn Help, 2024. <https://www.linkedin.com/help/linkedin/answer/a528190>.
- Lubin, Nathaniel, and Thomas Krendl Gilbert. “Accountability Infrastructure: How to Implement Limits on Platform Optimization to Protect Population Health,” June 12, 2023. https://www.platformaccountability.com/_files/ugd/424593_5af63341c56d48c4807afef21298f6f7.pdf.
- Lubin, Nathaniel, Yuning Liu, Amanda Yarnell, et al. “Social Media Harm Abatement: Mechanisms for Transparent Public Health Assessment.” *Annals of the New York Academy of Science* (Forthcoming), 2025.
- Lukoff, Kai, Ulrik Lyngs, Himanshu Zade, et al. “How the Design of YouTube Influences User Sense of Agency.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. CHI ’21, 2021. <https://doi.org/10.1145/3411764.3445467>.
- Maryland. “Maryland Kids Code,” 2024. <https://mgaleg.maryland.gov/2024RS/bills/sb/sb0571T.pdf>.
- Meta. “Instagram Feed AI System.” Transparency Center, 2024. <https://transparency.meta.com/features/explaining-ranking/ig-feed>.
- — —. “Meta Pixel: Measure, Optimize & Retarget Ads on Facebook & Instagram.” Meta for Business. Accessed January 28, 2025. <https://www.facebook.com/business/tools/meta-pixel>.
- — —. “Our Approach to Explaining Ranking.” Transparency Center, December 31, 2023. <https://transparency.meta.com/features/explaining-ranking>.
- — —. “Review Your Activity off Meta Technologies.” Facebook Help Center. Accessed January 28, 2025. <https://www.facebook.com/help/2207256696182627/>.
- — —. “Transparency Reports.” Transparency Center. Accessed February 4, 2025. <https://transparency.meta.com/reports/>.
- — —. “Widely Viewed Content Report: What People See on Facebook.” Transparency Center, Q3 2024. <https://transparency.meta.com/data/widely-viewed-content-report/>.
- Millecamp, Martijn, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. “Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces.” In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 101–9. UMAP ’18, 2018. <https://doi.org/10.1145/3209219.3209223>.

- Milli, Smitha, Luca Belli, and Moritz Hardt. "From Optimizing Engagement to Measuring Value." In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 714–22. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445933>.
- Milli, Smitha, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media." December 7, 2024. <https://doi.org/10.48550/arXiv.2305.16941>.
- Minnesota. "Prohibiting Social Media Manipulation Act." 2024. https://www.revisor.mn.gov/bills/text.php?number=HF4400&version=0&session=ls93&session_year=2024&session_number=0.
- — —. "SF 2716." 2023. <https://www.revisor.mn.gov/bills/bill.php?b=Senate&f=SF2716&ssn=0&y=2023>.
- Moehring, Alex. "Personalization, Engagement, and Content Quality on Social Media: An Evaluation of Reddit's News Feed." May 30, 2024. <https://doi.org/10.31219/osf.io/8yuwe>.
- National Academies of Sciences, Engineering, and Medicine. *Social Media and Adolescent Health*. National Academies Press, 2024. <https://doi.org/10.17226/27396>.
- New York. "Stop Addictive Feeds Exploitation (SAFE) for Kids Act," 2023. <https://www.nysenate.gov/legislation/bills/2023/S7694/amendment/A>.
- NewsGuard. "News Reliability Ratings." NewsGuard. Accessed December 20, 2024. <https://www.newsguardtech.com/solutions/news-reliability-ratings/>.
- North Carolina. "Social Media Algorithmic Control in IT Act," 2023. <https://webservices.ncleg.gov/ViewBillDocument/2023/3923/0/DRH10282-LRa-77C>.
- Ofcom. "Experiences of Using Online Services," 2024. <https://www.ofcom.org.uk/media-use-and-attitudes/online-habits/internet-users-experience-of-harm-online/>.
- — —. "Fewer than Half of Social Media Users Find Content Controls Effective." February 28, 2024. <https://www.ofcom.org.uk/media-use-and-attitudes/online-habits/fewer-than-half-of-social-media-users-find-content-controls-effective/>.
- Office of the Surgeon General. "Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory." Publications and Reports of the Surgeon General. US Department of Health and Human Services, 2023. <http://www.ncbi.nlm.nih.gov/books/NBK594761/>.
- Oklahoma. "Oklahoma Social Media Transparency Act." 2023. <http://www.oklegislature.gov/BillInfo.aspx?Bill=hb2548&Session=2400>.
- Ovadya, Aviv. "Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure." The Belfer Center for Science and International Affairs, 2021.

<https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure>.

Ovadya, Aviv, and Luke Thorburn. “Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance.” Knight First Amendment Institute at Columbia University, 2023. <http://knightcolumbia.org/content/bridging-systems>.

Papadamou, Kostantinos, Antonis Papasavva, Savvas Zannettou, et al. “Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children.” *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 26, 2020): 522–33. <https://doi.org/10.1609/icwsm.v14i1.7320>.

Park, Sora, Caroline Fisher, Terry Flew, and Uwe Dulleck. “Global Mistrust in News: The Impact of Social Media on Trust.” *International Journal on Media Management* 22, no. 2 (April 2, 2020): 83–96. <https://doi.org/10.1080/14241277.2020.1799794>.

Paruthi, Shalini, Lee J. Brooks, Carolyn D’Ambrosio, et al. “Consensus Statement of the American Academy of Sleep Medicine on the Recommended Amount of Sleep for Healthy Children: Methodology and Discussion.” *Journal of Clinical Sleep Medicine* 12, no. 11 (n.d.): 1549–61. <https://doi.org/10.5664/jcsm.6288>.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2016. <https://www.hup.harvard.edu/books/9780674970847>.

Perez, Sarah. “Reddit Locks down Its Public Data in New Content Policy, Says Use Now Requires a Contract.” *TechCrunch*, May 9, 2024. <https://techcrunch.com/2024/05/09/reddit-locks-down-its-public-data-in-new-content-policy-says-use-now-requires-a-contract/>.

Pershan, Claire, and Jesse McCrosky. “No Perfect Solution to Platform Profiling Under Digital Services Act.” *Tech Policy Press*, August 25, 2023. <https://techpolicy.press/no-perfect-solution-to-platform-profiling-under-digital-services-act>.

Peters, Dorian, Rafael A. Calvo, and Richard M. Ryan. “Designing for Motivation, Engagement and Wellbeing in Digital Experience.” *Frontiers in Psychology* 9 (May 28, 2018). <https://doi.org/10.3389/fpsyg.2018.00797>.

Peukert, Christian, Ananya Sen, and Jörg Claussen. “The Editor and the Algorithm: Recommendation Technology in Online News.” *Management Science* 70, no. 9 (September 2024): 5816–31. <https://doi.org/10.1287/mnsc.2023.4954>.

Piccardi, Tiziano, Martin Saveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne L. Tsai, and Michael Bernstein. “Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity.” November 22, 2024. <https://doi.org/10.48550/arXiv.2411.14652>.

- Pinterest Engineering. “How Holdout Groups Drive Sustainable Growth.” *Pinterest Engineering Blog*, February 17, 2017.
<https://medium.com/pinterest-engineering/how-holdout-groups-drive-sustainable-growth-35a4786c3801>.
- Pizzo Frey, Tracy, Jason Mills, Margaret Mitchell, et al. “Recommendation Systems in Social Media.” Common Sense Media, October 7, 2024.
<https://www.common Sense Media.org/ai-ratings/recommendation-systems-in-social-media>.
- Public Interest Tech Lab. “Evaluating News Feed Ranking Experiments.” FBarchive, 2022.
<https://fbarchive.org/doc/odoc893107>.
- Radesky, Jenny, Enrica Bridgewater, Shira Black, et al. “Algorithmic Content Recommendations on a Video-Sharing Platform Used by Children.” *JAMA Network Open*, no. 5 (May 29, 2024).
<https://doi.org/10.1001/jamanetworkopen.2024.13855>.
- Rafieian, Omid, and Hema Yoganarasimhan. “AI and Personalization.” *Artificial Intelligence in Marketing* 20 (March 13, 2023): 77–102. <https://doi.org/10.1108/S1548-643520230000020004>.
- Reddit. “About the Reddit Pixel.” Accessed January 28, 2025.
<https://business.reddithelp.com/s/article/reddit-pixel>.
- — —. “Settings.” Privacy. Accessed January 28, 2025. <https://www.reddit.com/settings/privacy>.
- Redmiles, Elissa M., Michelle L. Mazurek, and John P. Dickerson. “Dancing Pigs or Externalities? Measuring the Rationality of Security Decisions.” In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 215–32, 2018. <https://doi.org/10.1145/3219166.3219185>.
- Samson, Alain, and Benjamin G. Voyer. “Two Minds, Three Ways: Dual System and Dual Process Models in Consumer Psychology.” *AMS Review* 2, no. 2 (December 1, 2012): 48–71.
<https://doi.org/10.1007/s13162-012-0030-9>.
- Schnabel, Tobias, Saleema Amershi, Paul N. Bennett, Peter Bailey, and Thorsten Joachims. “The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation.” In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 991–1000. SIGIR ’20, 2020. <https://doi.org/10.1145/3397271.3401117>.
- Singh, Ashudeep, Yoni Halpern, Nithum Thain, et al. “Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach.” In *FACCTRec Workshop on Responsible Recommendation*, 2020. https://www.ashudeepsingh.com/publications/facctrec2020_singh_et_al.pdf.
- Skiera, AJ. “What Gen Z Thinks about Its Social Media and Smartphone Usage.” *The Harris Poll*, 2024.
<https://theharrispoll.com/briefs/gen-z-social-media-smart-phones/>.
- Smith, Ben. “How TikTok Reads Your Mind.” *The New York Times*, December 6, 2021.
<https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html>.

- Snap. “How We Rank Content on Discover.” Snapchat Support. Accessed December 20, 2024.
<https://help.snapchat.com/hc/en-us/articles/8961631424020-How-We-Rank-Content-on-Discover>.
- — —. “How We Rank Content on Spotlight.” Snapchat Support. Accessed December 20, 2024.
<https://help.snapchat.com/hc/en-us/articles/8961653169940-How-We-Rank-Content-on-Spotlight>.
- — —. “Snapchat Transparency Report.” Privacy, Safety, and Policy Hub, 2024.
<https://values.snap.com/privacy/transparency>.
- Somerville, Leah H. “The Teenage Brain: Sensitivity to Social Evaluation.” *Current Directions in Psychological Science*, no. 2 (April 1, 2013): 121–27. <https://doi.org/10.1177/0963721413476512>.
- Stokel-Walker, Chris. “Twitter’s \$42,000-per-Month API Prices Out Nearly Everyone.” *Wired*. Accessed January 15, 2025. <https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/>.
- Stray, Jonathan. “Dependent Variables: The Outcomes We Will Measure.” *The Prosocial Ranking Challenge*, February 2, 2024.
<https://rankingchallenge.substack.com/p/dependent-variables-the-outcomes>.
- — —. “Designing Recommender Systems to Depolarize.” July 11, 2021.
<https://doi.org/10.48550/arXiv.2107.04953>.
- Stray, Jonathan, Alon Halevy, Parisa Assar, et al. “Building Human Values into Recommender Systems: An Interdisciplinary Synthesis.” *ACM Transactions on Recommender Systems* 2, no. 3 (June 5, 2024): 1–57. <https://doi.org/10.1145/3632297>.
- Stray, Jonathan, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. “What Are You Optimizing for? Aligning Recommender Systems with Human Values.” July 22, 2021.
<https://doi.org/10.48550/arXiv.2107.10939>.
- Stroud, Natalie Jomini, Ashley Muddiman, and Joshua M Scacco. “Like, Recommend, or Respect? Altering Political Behavior in News Comment Sections.” *New Media & Society* 19, no. 11 (November 1, 2017): 1727–43. <https://doi.org/10.1177/1461444816642420>.
- Sullivan, David, and Jason Pielmeier. “Unpacking ‘Systemic Risk’ Under the EU’s Digital Service Act.” *Tech Policy Press*, July 19, 2023.
<https://techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act>.
- Superior Court of the State of California. “Neville v. Snap Inc.,” 2023.
<https://fingfx.thomsonreuters.com/gfx/legaldocs/xmpjrooypr/frankel-nevillevsnap--amendedcomplaint.pdf>.
- Supreme Court of the United States. “Gonzalez et al. v. Google,” 2023.
https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf.
- — —. “NetChoice v. Moody,” 2024. https://www.supremecourt.gov/opinions/23pdf/22-277_d18f.pdf.

- — —. “Taamneh et al. v. Twitter,” 2023.
https://www.supremecourt.gov/opinions/22pdf/21-1496_d18f.pdf.
- Texas. “HB20,” 2021. <https://capitol.texas.gov/BillLookup/BillStages.aspx?LegSess=872&Bill=HB20>.
- — —. “SCOPE Act,” 2023. <https://capitol.texas.gov/tlodocs/88R/billtext/html/HB00018F.htm>.
- The YouTube Team. “Testing New Ways to Offer Viewers More Context and Information on Videos.”
YouTube Official Blog, June 17, 2024.
<https://blog.youtube/news-and-events/new-ways-to-offer-viewers-more-context/>.
- Thiel, David, Renee DiResta, and Alex Stamos. “Addressing the Distribution of Illicit Sexual Content by Minors Online.” Stanford Cyber Policy Center, June 6, 2023.
<https://cyber.fsi.stanford.edu/news/addressing-distribution-illicit-sexual-content-minors-online>.
- Thorburn, Luke, Priyanjana Bengani, and Jonathan Stray. “How Platform Recommenders Work.”
Understanding Recommenders, November 23, 2022.
<https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a>.
- TikTok. “How TikTok Recommends Content.” TikTok Help Center. Accessed December 20, 2024.
<https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content>.
- — —. “Transparency Center,” 2024. <https://www.tiktok.com/transparency/en-us/>.
- Turow, Joseph. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. Yale University Press, 2012.
- United Nations Independent Investigative Mechanism for Myanmar. “Anti-Rohingya Hate Speech On Facebook: Content and Network Analysis.” March 2024.
https://iimm.un.org/wp-content/uploads/2024/03/Hate-Speech-Report_EN.pdf.
- United States. “Algorithmic Accountability Act,” September 21, 2023.
<https://www.congress.gov/bill/118th-congress/senate-bill/2892>.
- — —. “Algorithmic Justice and Online Platform Transparency Act,” July 13, 2023.
<https://www.congress.gov/bill/118th-congress/senate-bill/2325>.
- — —. “DISCOURSE Act,” March 22, 2023.
<https://www.congress.gov/bill/118th-congress/senate-bill/921/actions>.
- — —. “Health Misinformation Act,” July 22, 2021.
<https://www.congress.gov/bill/117th-congress/senate-bill/2448>.
- — —. “Justice Against Malicious Algorithms Act,” October 15, 2021.
<https://www.congress.gov/bill/117th-congress/house-bill/5596#>.
- — —. “Kids Online Safety and Privacy Act,” June 21, 2023.
<https://www.congress.gov/bill/118th-congress/senate-bill/2073/text>.
- — —. “Platform Accountability and Transparency Act,” June 8, 2023.
<https://www.congress.gov/bill/118th-congress/senate-bill/1876>.

— — —. “Platform Integrity Act,” December 27, 2022.

<https://www.congress.gov/bill/117th-congress/house-bill/9695>.

— — —. “Preventing the Algorithmic Facilitation of Rental Housing Cartels Act,” January 30, 2024.

<https://www.congress.gov/bill/118th-congress/senate-bill/3692>.

— — —. “Protecting Americans from Dangerous Algorithms Act,” March 23, 2021.

<https://www.congress.gov/bill/117th-congress/house-bill/2154#>.

United States Court of Appeals for the Ninth Circuit. “Dyroff v. The Ultimate Software Group,” 2019.

<https://cdn.ca9.uscourts.gov/datastore/opinions/2019/08/20/18-15175.pdf>.

— — —. “Lemmon v. Snap,” 2021.

<https://cdn.ca9.uscourts.gov/datastore/opinions/2021/05/04/20-55295.pdf>.

— — —. “X Corp. v. Bonta,” 2024.

<https://cdn.ca9.uscourts.gov/datastore/opinions/2024/09/04/24-271.pdf>.

United States Court of Appeals for the Second Circuit. “Force v. Facebook,” 2019.

<https://www.govinfo.gov/content/pkg/USCOURTS-ca2-18-00397/pdf/USCOURTS-ca2-18-00397-0.pdf>.

United States Court of Appeals for the Third Circuit. “Anderson v. TikTok,” 2024.

<https://www2.ca3.uscourts.gov/opinarch/223061p.pdf>.

United States District Court for the District of Utah. “NetChoice v. Reyes,” 2024.

https://netchoice.org/wp-content/uploads/2023/12/NetChoice-v-Reyes_Official-Complaint_FINAL-Filed.pdf.

United States District Court for the Northern District of California. “In Re: Social Media Adolescent Addiction/Personal Injury Products Liability Litigation.” Accessed February 13, 2025.

<https://cand.uscourts.gov/in-re-social-media-adolescent-addiction-personal-injury-products-liability-litigation-mdl-no-3047/>.

— — —. “NetChoice v. Bonta,” 2024.

<https://oag.ca.gov/system/files/attachments/press-docs/Order%20on%20Preliminary%20Injunction.pdf>.

United States District Court for the Western District of Texas. “Computer & Communications Industry Association and NetChoice v. Paxton,” 2024.

https://netchoice.org/wp-content/uploads/2024/07/HB-18-Complaint_As-Filed.pdf.

Utah. “S.B. 194 Social Media Regulation Amendments,” 2024.

<https://le.utah.gov/~2024/bills/static/SB0194.html>.

Vermont. “S.289,” 2024. <https://legislature.vermont.gov/bill/status/2024/S.289>.

Virginia. “Consumer Data Protection Act; Social Media Platforms,” 2024.

<https://legacylis.virginia.gov/cgi-bin/legp604.exe?241+sum+HB1115>.

- Washington Post. “The Washington Post Launches a New Commenting Experience Exclusively for Subscribers.” *WashPost PR Blog*, December 16, 2024.
<https://www.washingtonpost.com/pr/2024/12/16/washington-post-launches-new-commenting-experience-exclusively-subscribers/>.
- Wirtschafter, Valerie, and Sharanya Majumder. “Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter’s Community Notes.” *Journal of Online Trust and Safety* 2, no. 1 (September 21, 2023). <https://doi.org/10.54501/jots.v2i1.139>.
- Wojcik, Stefan, Sophie Hilgard, Nick Judd, et al. “Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation,” 2022.
<https://doi.org/10.48550/ARXIV.2210.15723>.
- Wu, Tim. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. Vintage, 2016.
- X. “Twitter’s Recommendation Algorithm.” *X Engineering*, March 31, 2023.
https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.
- — —. “X Transparency Center,” 2024. <https://transparency.x.com/en>.
- Zhang, Mingrui Ray, Kai Lukoff, Raveena Rao, Amanda Baughan, and Alexis Hiniker. “Monitoring Screen Time or Redesigning It? Two Approaches to Supporting Intentional Social Media Use.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. CHI ’22, 2022. <https://doi.org/10.1145/3491102.3517722>.
- Zuckerberg, Mark. “A Blueprint for Content Governance and Enforcement,” 2018.
https://www.facebook.com/notes/751449002072082/?checkpoint_src=any.