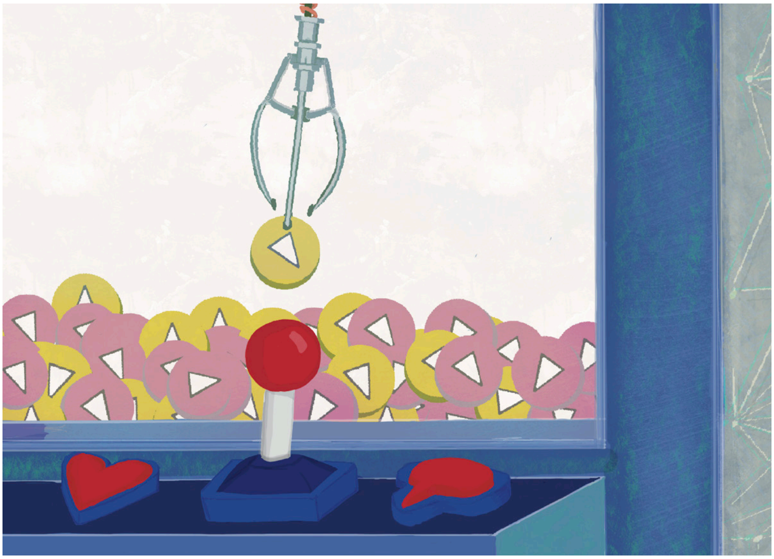
OPTIMIZING FOR WHAT?
ALGORITHMIC AMPLIFICATION AND SOCIETY



Understanding Social Media Recommendation Algorithms

By Arvind Narayanan

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

This essay was written by Arvind Narayanan, professor of computer science at Princeton University, during his time as a visiting senior research scientist at the Knight First Amendment Institute. Prof. Narayanan oversaw a year-long project studying how social media algorithms amplify some speech and suppress others. In partnership with the Knight Institute, Prof. Narayanan organized a major symposium, “Optimizing for What? Algorithmic Amplification and Society.” The essays written for that symposium, and additional essays from Prof. Narayanan on the same theme, are forthcoming from the Knight Institute.

The full series is available at knightcolumbia.org/research/

CONTENTS

ABSTRACT

Recommendation algorithms mediate online speech and thus shape society. In this essay, I start by explaining the basics of information propagation on social media. Then I turn to the role of recommendation algorithms and engagement optimization—a role that has become more and more prominent over the last decade. I hope to refute the idea that recommendation algorithms are hard to understand because they use sophisticated mathematical techniques and reveal what is truly hard about understanding their societal effects. Finally, I examine the problematic normative assumptions behind today's recommendation algorithms.

INTRODUCTION

When we speak online—when we share a thought, write an essay, post a photo or video—who will hear us? The answer is determined in large part by algorithms. In computer science, the algorithms driving social media are called recommender systems. These algorithms are the engine that makes Facebook and YouTube what they are, with TikTok more recently showing the power of an almost purely algorithm-driven platform.

In debates about the effects of social media, discussion of algorithms tends to be superficial. They are often assumed to be black boxes that are too complicated to understand. This is unfortunate. In fact, there is a lot that is known about how these algorithms operate. But this knowledge is not yet broadly accessible.

I think a broader understanding of recommendation algorithms is sorely needed. Policymakers and legal scholars must understand these algorithms so that they can sharpen their thinking on platform governance; journalists must understand them so that they can explain them to readers and better hold platforms accountable; technologists must understand them so that the platforms of tomorrow may be better than the ones we have; researchers must understand them so that they can get at the intricate interplay between algorithms and human behavior. Content creators would also benefit from understanding them so that they can better navigate the new landscape of algorithmic distribution. More generally, anyone concerned about the impact of algorithmic platforms on themselves or on society may find this essay of interest.

I hope to show you that social media algorithms are simple to understand. In addition to the mathematical principles of information cascades (which are independent of any platform), it's also straightforward to understand what recommendation algorithms are trained to do, and what inputs they use. Of course, companies' lack of transparency about some of the details is a big problem, but that's a separate issue from the details being hard to understand—they aren't. In this regard, recommendation algorithms are like any other technology, say a car or a smartphone. Many details of those products are proprietary, but we can and do understand how cars and smartphones work. Once we understand the basics of recommendation algorithms, we can also gain clarity on *which* details matter for transparency.

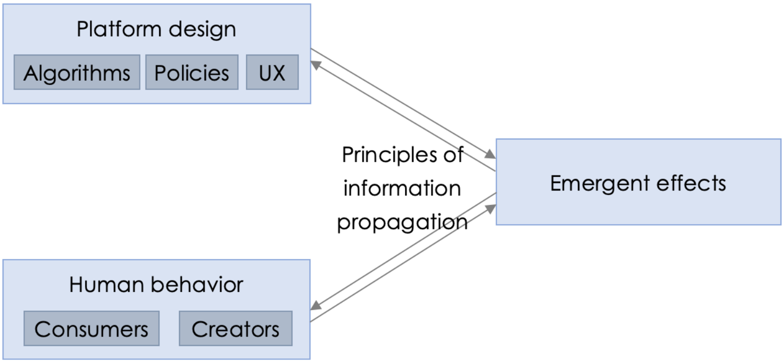
In composing this essay, I've relied on the computer science literature on social networks, recommender systems, and related topics; companies' (minimal) public documentation of their algorithms; the documents leaked by Facebook whistleblower Frances Haugen;[1] and a few of my own observations. My contribution is to synthesize this information, introduce conceptual frameworks for understanding it, and describe it without jargon. My goal is not to explain the tech for its own sake but rather with a view to understanding its societal effects. To that end, I've also included commentary on problems with algorithmic recommendations, specifically algorithms that optimize for engagement.

SOCIAL MEDIA PLATFORMS ARE "COMPLEX SYSTEMS"

A COMPLEX SYSTEM is one whose behavior arises in nonlinear, often unpredictable ways from those of its parts.[2] This phenomenon is called emergence. For example, traffic is famously a complex system.[3] Adding a road to a network of roads, keeping everything else the same, can *slow* the overall traffic through it.

Social media platforms are complex systems subject to various emergent behaviors and patterns of feedback. Social movements can form in a flash as attention to an event or a cause begets more attention.[4] U.S. politicians learned to be less civil because such posts garnered more attention.[5] Matias and Wright document many other feedback loops.[6]

Figure 1: The effects of information propagation on platforms emerge through the interaction of design and user behavior, based on underlying mathematical principles. Design comprises algorithms, the user interface, and various policies, such as content moderation policies. Platform designers, users, and content creators all adapt to emergent effects.



Platform design matters but isn't the whole picture. Many pathologies of social media are attributed either to human behavior or to the algorithms that mediate information propagation when they are in fact the result of both. Consider these examples of *either-or* thinking to explain observed or hypothesized phenomena:

- “People on Twitter are too negative,” versus “The Twitter algorithm rewards negativity.”
- “YouTube’s algorithm pushes users into rabbit holes,” versus “It’s not the algorithm, it’s users’ natural behavior.”

My view is that these and many other phenomena are emergent effects of human-algorithm interactions. The research community is not close to being able to fully observe and explain the underlying feedback loops, both because the methods remain immature and because of lack of adequate access to platforms.

THERE ARE MANY DIFFERENT SOCIAL MEDIA ALGORITHMS

THERE ARE MANY algorithms behind any large social media platform. The table shows a rough categorization of the major algorithms. One set of algorithms *processes* content. Another set of algorithms *propagates* it, that is, helps determine who sees what. The set of content processing algorithms is relatively fluid as new types of content become prominent and new algorithmic capabilities emerge. The set of content propagation algorithms is relatively stable.[7]

Table 1: Major social media algorithms.

Content processing	**Content propagation**
Face recognition	Search
Image filters	Recommendation (feeds)
Annotation (e.g., image tagging)	Ad delivery and targeting
Audio transcription	Content moderation
Language translation	Friend recommendation
Augmented & virtual reality	Notification
…	Trending
	…

While all these algorithms are important, my main focus in this essay is on content recommendation algorithms. These algorithms generate our social media feeds. They show up in a few other places, like YouTube sidebar recommendations. They aren't limited to social media or user-generated content: Movie recommendations on Netflix and product recommendations on Amazon belong to the same class of algorithms. Why focus on recommendation algorithms? Compared to search, recommendation drives a bigger (and increasing) fraction of engagement. More importantly, the platform has almost complete control over what to recommend a user, whereas search results are relatively tightly constrained by the search term.

Even the "recommendation algorithm" on any large platform is in fact a whole suite of algorithms, but they are tightly coupled, so I will refer to them collectively as "the algorithm." Sometimes I refer to recommendation algorithms collectively, and sometimes I refer to a specific platform's algorithm.

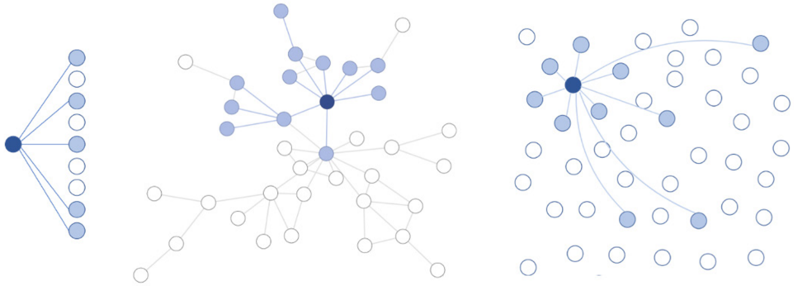
THREE TYPES OF INFORMATION PROPAGATION: SUBSCRIPTION, NETWORK, AND ALGORITHM

NOT ALL SOCIAL MEDIA feeds are algorithmic, and not all the emergent effects we're concerned with involve algorithms. It's extremely helpful to understand the three fundamental ways in which the information-propagation component of a platform can be designed. No platform follows precisely one of these models; they all mix and match. Still, it's best to understand the basic models first, and then think about how they are combined in any given system.

Table 2: Three stylized models of information propagation.

	Subscription	**Network**	**Algorithm**
What a user sees	Posts by those they've subscribed to	Posts by (or shared by) those they've subscribed to	Posts the algorithm predicts the user will like best
Examples	Newspapers, Substack, FB pre-2009, IG pre-2022	Word of mouth, the web, Twitter pre-2016, Mastodon	TikTok, Google Discover, YouTube
What impacts a post's reach	Poster's subscriber count	Both subscriber count and content	The content of the post

Figure 2: Three models of information propagation: subscription, network, and algorithm, showing the propagation of one individual post. In the subscription model, the post reaches those who have subscribed to the poster. In the network model, it cascades through the network as long as users who see it choose to further propagate it. In the algorithmic model shown here, users with similar interests (as learned by the algorithm based on their past engagement) are depicted closer to each other. The more similar a user's interests are to the poster's, the more likely they are to be recommended the post. Of course, other algorithmic logics are possible.



The subscription model is straightforward: Each user subscribes to a set of creators, and their feed consists of posts from their creators. In traditional media, we call this broadcast. If you subscribe to a set of newspapers, or a set of cable channels, then that's the content you receive.

Note that originally (in the 2000s), neither Facebook nor Twitter had the ability to reshare or retweet posts in your feed. This critical feature is what separates the subscription model from the network model. In the network model, a user sees not only posts created by those they've subscribed to, but also posts that those users choose to amplify, creating the possibility of information cascades ("viral" posts). Before Twitter introduced the algorithmically ranked feed in 2016, it followed a network model almost purely.[8] This is usually what people mean by "chronological feed."

Let's take a minute to understand the algorithmic model. Very few platforms implement a purely algorithmic model, so it's tricky to get a good intuition for it. In this model, the posts a user sees are those that the algorithm predicts they are most likely to engage with (the definition of engagement is critical, but let's put that aside for now). *There is no social network*. That is, there is no ability for users to follow, subscribe to, or connect with others—or, if there is, it doesn't determine what shows up on a user's feed.

TikTok's "For You Page," which is where users spend almost all of their time, is famously algorithmic.[9] Google has an algorithmic news recommendation product called Google Discover. Surprisingly little has been said about it given that it is a product that Google heavily promotes to its over 3 billion mobile users.[10] YouTube uses a mix of the subscription and algorithmic models (without much in the way of network dynamics) but heavily tilted toward algorithms.[11]

Over the past two decades, the progression has been from the subscription model to the network model to the algorithmic model. We appear to be in the middle of the latter shift (from network to algorithm), notably with Instagram and Facebook. Other platforms are facing similar pressure as well, because of the success of TikTok. Any such shift has major impacts on the platform as a business, on the type of content that's amplified, and on the user experience. For example, Instagram's changes led to a user outcry that forced it to roll back some changes.[12]

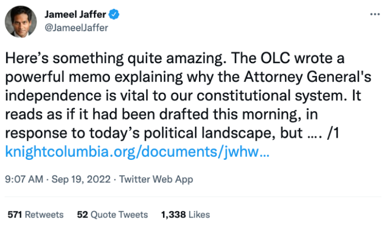
Perhaps the biggest impact of the shift to the algorithmic model is on content creators. In the subscription and network models, creators can focus on building their network. In the algorithmic model, that doesn't help, because the number of subscribers is irrelevant to how posts will perform. (If this sounds unintuitive, it's because no platform implements a purely algorithmic model, and the network always matters to some degree.) Instead, the audience for each post is independently optimized based on the topic and the "quality" of the post. In this idealized setting, considering other factors such as the performance of past posts by that creator can only detract from the goal of optimizing the audience for the present post. Of course, the algorithm's notion of quality might not be normatively desirable: The content that it amplifies might not align with our idea of healthy discourse. In any case, the less emphasis there is on the network, the less predictability and control creators have over the reach of their content. An algorithm change that devalues a particular type of content could wipe out a creator at any time.

To reiterate: The three models I've presented are idealized, and I've found the categorization helpful as an analytical lens, but almost no real platform adheres entirely to any one model. For example, even platforms that implement the subscription and network models tend to use recommendation algorithms in one important way: to rank posts in a user's feed, although not to determine which posts to include or exclude. Most users don't consume their entire feed: For example, Instagram reported that in 2016, users saw only 30% of the posts in their feed.[13] This means that the ranking algorithm makes a big difference to engagement. So most of what I'll say in this essay about the algorithmic model applies broadly to social media, not just to the platforms I've categorized as algorithmic.

The three models increase in complexity with respect to the way information propagates. The subscription model is straightforward, so I won't say much more about it. But there's a lot to say about the network model, so I'll discuss that in the next few sections. Understanding those details will help us better appreciate the significance of the turn to algorithms.

NETWORKS ENABLE VIRALITY

CONSIDER THESE TWO TWEETS: One is an in-depth thread about an intriguing document, and the other regurgitates political talking points.[14] One of these tweets was viral, and the other wasn't. Which is which?

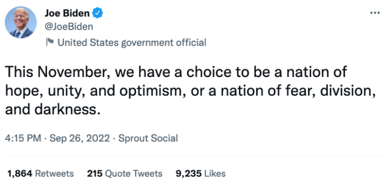


Jameel Jaffer
@JameelJaffer

Here's something quite amazing. The OLC wrote a powerful memo explaining why the Attorney General's independence is vital to our constitutional system. It reads as if it had been drafted this morning, in response to today's political landscape, but …. /1
knightcolumbia.org/documents/jwhw…

9:07 AM · Sep 19, 2022 · Twitter Web App

571 Retweets 52 Quote Tweets 1,338 Likes



Joe Biden
@JoeBiden
United States government official

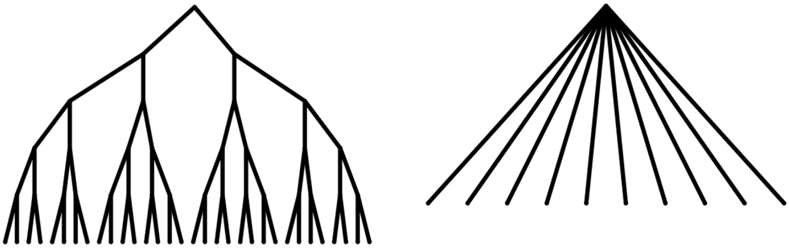
This November, we have a choice to be a nation of hope, unity, and optimism, or a nation of fear, division, and darkness.

4:15 PM · Sep 26, 2022 · Sprout Social

1,864 Retweets 215 Quote Tweets 9,235 Likes

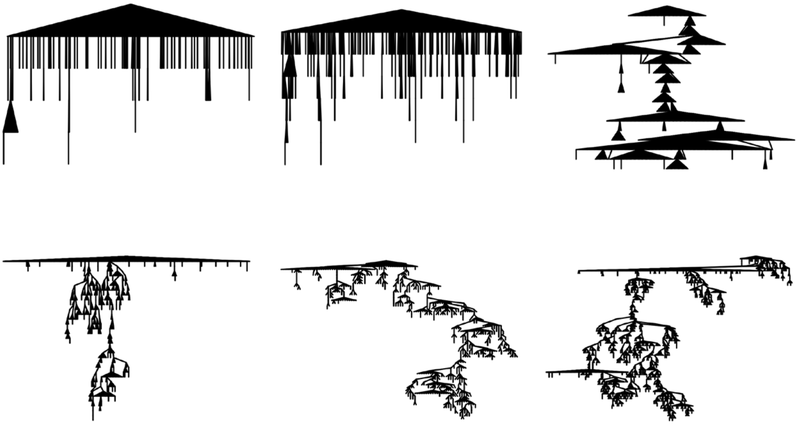
Based on the retweet and like counts, @JoeBiden's tweet was more popular. But virality is not popularity. It's about whether the piece of content spread in the manner of a virus, that is, from person to person in the network, rather than as a broadcast. It turns out that this can be measured, and we can assign a single number called structural virality that captures how viral something is.

Figure 3: Information cascade patterns representing viral and broadcast propagation (stylized). From Goel et al.[15]



Structural virality is the answer to the question: "How far from the poster did the post travel through the network?" It's a simple question that reveals a lot, illustrated by the stylized trees (in computer science, "trees" are drawn upside down). The cascade pattern of a tweet like @JameelJaffer's would look like the one on the left, retweeted by many people who aren't following the original account, whereas @JoeBiden's tweet would look like the one on the right. The structural virality of a post is the number of degrees of separation, on average, between users in the corresponding tree. The deeper the tree, with more branches, the greater the structural virality.

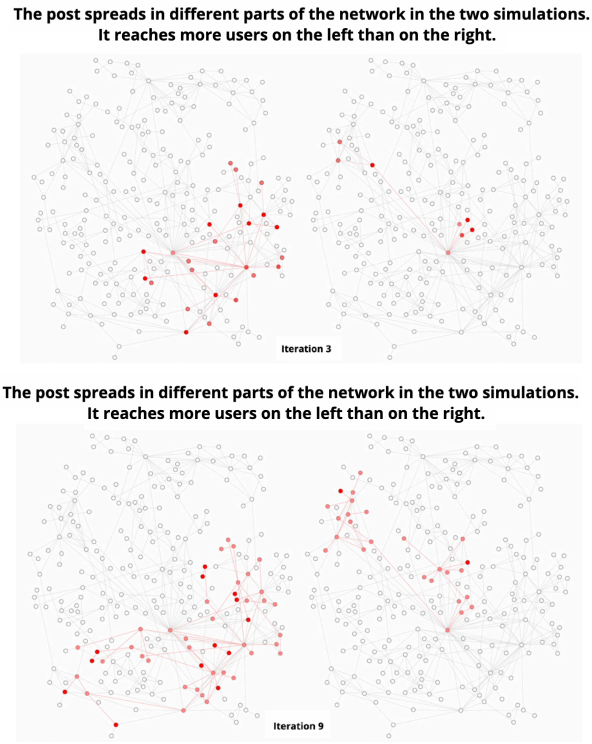
Structural virality was defined in a paper by Goel, Anderson, Hofman, and Watts.[16] To illustrate, they show six actual Twitter cascades with varying degrees of virality, ordered from least to most viral.



VIRALITY IS UNPREDICTABLE

THE FIGURE SHOWS a visualization of how information spreads in a social network. Both simulations use the same network with nodes (users) behaving identically: resharing information that they see with a certain probability. Purely due to this randomness, the information cascade evolves very differently in the two simulations. Not only does the cascade reach a much greater number of nodes in one simulation than the other, it also spreads through a different part of the network.

Figure 4: Simulation of information cascades in a social network, illustrating the unpredictability of virality.


The post spreads in different parts of the network in the two simulations.
It reaches more users on the left than on the right.
Iteration 3
The post spreads in different parts of the network in the two simulations.
It reaches more users on the left than on the right.
Iteration 9

Note that the unpredictability of user behavior is inevitable. Simply depending on the time of day that a user happens to be on the app, the set of posts they would see in their feed might differ substantially.

Research on real-world social networks supports the hypothesis that reach is unpredictable. A 2016 study attempted to predict the number of retweets of a given tweet based on the information available when it was tweeted: the content of the tweet and information about the creator.[17] The most accurate model in the study could explain no more than half the variance in retweet counts. More significantly, it was hardly more accurate than a model that ignored tweet content and was restricted to only looking at user information (follower count, performance of past tweets, etc.). Of course, for a given creator, the user information is fixed, and only the tweet content varies, so reach is essentially completely unpredictable, at least based on the methods used in the paper.

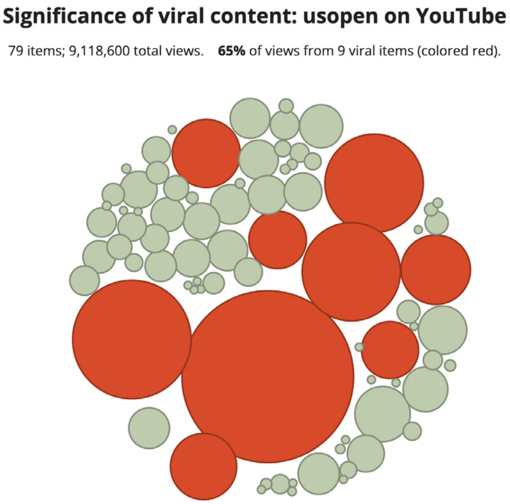
VIRAL CONTENT DOMINATES OUR ATTENTION

THE UNPREDICTABILITY OF VIRALITY is a fact of life for creators. It is made worse by the fact that only a small fraction of posts are likely to go viral. The structural virality paper quantifies this (on a global level rather than a per-creator level): In their dataset, less than 1 in 100,000 tweets is retweeted 1,000 times. Intuitively, this makes sense: Attention is finite, so there can be only a certain amount of viral content going around at any given time, and competition for popularity is intense.

My hypothesis is that on every major platform, for most creators, the majority of engagement comes from a small fraction of viral content. The data that I've seen from studies and from my own investigations is consistent with this: The distribution of engagement is highly skewed. A 2022 paper quantified this for TikTok and YouTube: On TikTok, the top 20% of an account's videos get 76% of the views, and an account's most viewed video is on average 64 times more popular than its median video.[18] On YouTube, the top 20% of an account's videos get 73% of the views, and an account's most viewed video is on average 40 times more popular than its median video. In general, the more significant the role of the algorithm in propagating content, as opposed to subscriptions or the network, the greater this inequality seems to be.

Here's a visualization of the significance of virality. For the purposes of this visualization, I define a viral post by a creator as one whose engagement is over five times the median engagement of that creator's posts. I use this alternative definition since structural virality is not publicly visible.[19] In reality, viral content is even more significant than appears from this kind of illustration, because virality is the main way to reach new audiences and gradually grow one's reach over time.

Figure 5: The significance of virality for one selected account. The level of skew shown here is quite common, though there is substantial variation between accounts.[20]

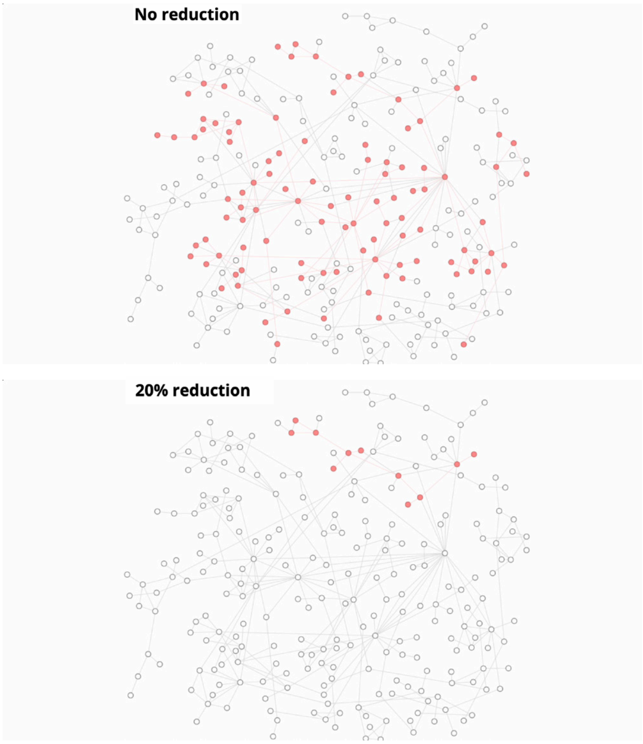


VIRAL CONTENT IS HIGHLY AMENABLE TO DEMOTION

DEMOTION, DOWNRANKING, reduction, or suppression, often colloquially called shadowbanning, is a "soft" content moderation technique in which content deemed problematic is shown to fewer users, but not removed from the platform.[21] There are many ways to implement it. For example, Facebook ranks demoted posts lower in users' feeds than they would otherwise rank, the idea being that users are less likely to encounter and further spread them.

A seemingly small interference by the platform can drastically decrease the reach of downranked content. To illustrate this, I use a simplified model of demotion and simulate varying degrees of demotion. Specifically, in this model, the post is demoted in such a way that the probability of a user seeing it (conditional on its appearing in their feed) decreases by 10%, 20%, or 30% respectively.

Figure 6: Simulation to illustrate the effect of demotion.



Without demotion, the post would reach the majority of the network. A 10% reduction has little impact; the reach remains almost the same. But a 20% reduction causes its reach to drop *tenfold*, and the content only reaches the poster's immediate network. The specific numbers here are not important; the point is that the effect of demotion on reach can be unpredictable, nonlinear, and sometimes drastic.

Demotion is nontransparent because it isn't necessarily noticeable by the poster's followers (as the post still appears in their feeds) and because low reach isn't automatically suspicious, since there is a large amount of variation in the natural reach of a poster's content. By the same token, users may sometimes incorrectly conclude that they have been "shadowbanned" when their reach is low.

THE CORE OF THE ALGORITHM IS ENGAGEMENT PREDICTION

PLATFORM COMPANIES may have many high-level goals they care about: ad revenue, keeping users happy and getting them to come back, and perhaps also less mercenary, more civic goals. But there's a problem: None of those goals are of much use when an algorithm is faced with a decision about what to feed a specific user at a specific moment in time. There isn't a good way to measurably connect this kind of micro-level decision to its long-term impact.

That's where engagement comes in. By engagement I mean any score that is defined only in terms of the moment-to-moment actions of the user. And that is its great virtue. For every single post in the user's feed, the algorithm receives feedback about whether and how the user engaged with it. That is why the primary objective of almost every recommendation algorithm on social media platforms is to rank the available content according to how likely it is that the user in question will engage with it.

In a sense, engagement is a proxy for high-level goals. A user who is engaged is more likely to keep returning and generate ad revenue for the platform. Because it is only a proxy, and developers are aware of its limits, there are many other considerations that go into platform algorithms. In terms of code, the part that calculates engagement may be only a small fraction. Yet it is the core logic, and it is a fruitful way to understand how content propagates on major platforms.

Here are some stylized examples of the flavors of engagement that various platforms optimize for.[22] A few caveats: I only list the *primary* optimization objective, which I think helps understand the essence of each platform. There may be many little tweaks in how engagement is calculated. This list reflects my best understanding based on the sources I cite. I have no insight into the matter beyond what has been publicly reported. In general, optimization targets are weighted averages of engagement signals available to the platform.

- Facebook optimizes for "Meaningful Social Interactions," a weighted average of Likes, Reactions, Reshares, and Comments.[23]
- Twitter, similarly, combines all the types of interaction that a user might have with a tweet.[24]
- YouTube optimizes for expected watch time, that is, how long the algorithm predicts the video will be watched.[25] If a user sees a video in their recommendations and doesn't click on it, the watch time is zero. If they click on it and hit the back button after a minute, the watch time is one minute. Before 2012, YouTube optimized for click-through rate instead, which led to clickbait thumbnails (such a sexualized imagery) becoming ubiquitous; hence the shift to watch time.[26]
- Less is known about TikTok's algorithm than those of the other major platforms, but it appears broadly similar: a combination of liking, commenting, and play time.[27] The documentation says that whether a video was watched to completion is a strong signal, and this factor probably gives the platform some of its uniqueness.[28] One indirect indication of the importance of this signal is that very short videos under 15 seconds—which are more likely to be played to completion, and thus score highly—continue to dominate the platform, despite the length restriction having been removed.[29] That might be because shorter videos are more likely to be watched to completion, and thus amplified and incentivized by the algorithm.
- Netflix originally optimized for suggesting movies that the user is likely to rate highly on a scale of one to five; this was the basis for a $1M recommender system competition, the Netflix Prize, in 2006.[30] But now it uses a more complex approach.[31]

Given a user and a post, the engagement prediction algorithm calculates a guess for how likely the user is to engage with the post if shown in their feed. To a first approximation, generating the user's feed is a matter of ranking all the posts that can be shown (in the order of decreasing predicted engagement). So Facebook would start with the post which it thinks you are most likely to like, react, reshare, or comment on; YouTube would start with the video that you are most likely to click on and continue watching, with videos that you are likely to watch for longer scoring higher. In Facebook's case, the set of candidate posts primarily consists of updates related to your friends or pages you follow, but this appears to be changing.[32] In YouTube's case, any video can potentially be recommended.

On top of this baseline logic, there are a whole bunch of secondary considerations.

- Keeping the computation tractable is an overriding consideration; slowing down the user experience is not considered an option. This is handled by first applying a candidate generation step that whittles the universe of content down to about a few hundred candidates.[33] It doesn't have to be accurate and only needs to select posts, not rank them, which is much faster. The engagement prediction/ranking algorithm is applied only to this smaller set. Once engagement predictions are calculated, the remaining considerations on this list are applied.
- If the user engages with content from a particular poster, each post from that poster will tend to rank highly. So the naive algorithm above would generate a feed that is overwhelmed by one or a small number of posters, which is undesirable. It is better to diversify the feed in terms of posters and topics. A diverse menu is also a defense against the algorithm's uncertainty about what the user wants at any given moment, because even the best algorithm is far from perfect at predicting engagement.
- That said, it is possible to tailor recommendations based on the user's "context": their geolocation, device, the content they have interacted with immediately prior, and so on. This context is one input to the engagement prediction algorithm.
- Platforms like Netflix and Spotify have found that explaining why a

recommendation was made makes them more persuasive.[34] They have made various modifications to their algorithms to enable this. Almost all platforms provide some degree of explanation, even if it isn't as central to the user experience as it is on Netflix or Spotify.

- Lately, platforms have started incorporating considerations of fairness to creators, such as gender fairness, to combat user biases and the way that algorithms amplify those biases if there is no intervention.[35]
- There's a trade-off between recommending content similar to what the user has engaged with in the past, which is a safe choice, and recommending new types of content so that the algorithm can learn whether the user is interested in it—and perhaps influence the user to acquire new interests. There's a class of algorithms devoted to optimally navigating this trade-off. TikTok is notable for its emphasis on exploration.[36]
- There is a near-endless list of subtle technical challenges. One example: If a user engaged with the first, third, and sixth posts in their feed, out of 10 posts, to what extent does that reflect the user's true preferences, versus the fact that people are generally more likely to pay more attention to posts closer to the top of their feed? The algorithm needs to disentangle these two factors.

While there are many differences in the particulars, the similarities between different platforms' recommendation algorithms overwhelm their differences. And the differences that do exist are generally specific to the design of the platforms. For example, YouTube optimizes for expected watch time, but Twitter doesn't, because Twitter is not video based. Spotify has the somewhat unique challenge of generating playlists that are coherent as a whole, rather than merely compiling a list of individually appealing track recommendations, so its logic departs somewhat substantially from the above. Perhaps for this reason, it relies more on content analysis and less on behavior.[37]

In other words, there is no competitive risk to platform companies from being more open about their algorithms. This might contradict one's mental picture of the algorithm being closely guarded secret sauce. In a blog post analyzing TikTok, I argued that this view is a myth, but that argument applies to other platforms too.[38]

In fact, most major platform companies are quite open about discussing their recommendation algorithms at academic and industry conferences and learn from each other. Much of what I wrote above is disclosed in well-known research papers. But it turns out that the details that matter from a research and engineering perspective are subtly different from those that matter to users and to society. And companies seem to have little incentive to be transparent about the algorithm with those stakeholders. That explains the current unsatisfactory and somewhat paradoxical state of algorithmic transparency. Besides, companies have shared precious little about the *effects* of algorithms. There have only ever been two published studies from major platform companies looking at the effects of their algorithms, as far as I'm aware.[39]

HOW TO PREDICT ENGAGEMENT

TO RECAP, THE RECOMMENDATION algorithms behind the major platforms are more similar than they are different in what they seek to accomplish. What varies quite a bit is *how* they optimize engagement: the signals they use and the computational techniques involved. But even here, the high-level logic is more-or-less the same. To predict engagement by a given user on a given post, most major recommendation algorithms try to answer the question:

> How did users similar to this user engage with posts similar to this post?

The intuition behind this logic is straightforward: Two people who have something in common—a hometown, a hobby, a community they are embedded in, a celebrity they follow—will both engage with posts that relate to that shared interests. While the intuition is compelling, the reason this approach is popular is that it has repeatedly proven to work well in practice.

To break it down, let's start with similarity between users. There are three main types of signals that are available: network, behavior, and demographics. Network refers to the user's interaction with others: following, subscription, commenting, and so on. Platforms vary in how much emphasis they place on this signal. Twitter relies heavily on the user-user network.[40] But on TikTok or YouTube, which place less emphasis on following, this signal is likely to be less useful. On platforms that don't have a social network, such as Netflix, the signal isn't available at all.

Behavior is the most critical signal. Two users are similar if they have engaged with a similar set of posts. Its importance is a matter of sheer volume. Here's a simple calculation: If a user spends an hour a day on TikTok for four years,[41] the average video length is 20 seconds, and they skip half the videos, the platform has interaction records on over *half a million* videos for that single user.

Demographics refers to attributes such as age, gender, and, more importantly, language and geography. Demographic information is useful when a user first joins the platform since there is little else to rely on. But once the user starts leaving a behavioral record, its importance rapidly diminishes.

In fact, the algorithm can automatically infer demographics like age, gender, and race as a side-effect of looking for patterns in the data, even if there was no intention to infer them. A few years ago, many Netflix users complained that their thumbnails seemed to be personalized to their race, such as Black users being shown a movie thumbnail containing two minor characters who are Black.[42] In response, the company pointed out that it doesn't ask users for race or ethnicity. But there is no contradiction: Demographic targeting can be an emergent effect of personalizing by viewing history. Race is a social construct, but it is one that is reflected in our behavior to some degree, and that is enough for the algorithm to reconstruct the category of race or something similar to it from behavioral records.

Turning from similarity between users to similarity between posts, the most obvious attribute that could be used for computing post similarity is content. The term content in this context usually refers to metadata (say, the title and description of a video) and less commonly the full content (i.e., the byte stream). The idea is simple: If a user likes a video on a particular topic, they will probably like other videos on the same topic.

To analyze content in this way, a set of "feature extraction" algorithms preprocesses posts and represents them in a form that's more digestible to algorithms: as a series of attributes (features). A simple example of a feature is the language or languages that appear in a post. Other features may be much more complex.

More and more, normative evaluations of posts are also included among the features. A canonical example is a score representing the likelihood that a post is hate speech. Posts may get algorithmically demoted based on such features—that is, their reach will be limited. This blurs the line between content moderation and algorithmic recommendation.

The other main signal relating to posts is, again, behavior: Two posts are similar if a similar set of users have engaged with them. Most platforms use both types of signals. As before, when a post has just been shared, the content signal predominates in importance, but as it accumulates an interaction history, behavior becomes more important.[43]

The most important fact to keep in mind is that the behavioral record is the fuel of the recommendation engine.

It might be surprising that recommendation algorithms are so simple to describe, given that large teams of highly skilled engineers work on them. But it takes a lot of ingenuity to translate high-level ideas of the sort I've described into an algorithm. In particular, keeping the computation tractable is a major challenge. The volume of information is vast: Based on the back-of-the-envelope calculations for TikTok above, the number of behavioral records may be of the order of a quadrillion (10^{15}). A naive algorithm—for instance, one that attempted to compute the affinity between each user and each post—would be millions of times slower than an optimized one, and no amount of hardware power can make up the difference.

A BRIEF HISTORY OF RECOMMENDATION ALGORITHMS

THE FIRST LARGE-SCALE and widely known online recommendation algorithm was deployed by Amazon in the late 1990s. Netflix followed soon after in 2000.[44] Both platforms quickly found that recommendations drove a substantial percentage of their purchases or rentals, and other companies began to adopt them. Considering their success in the e-commerce sector, it's surprising that social media platforms took so long to make recommendation algorithms a key part of how they work: It only started happening in the 2010s.

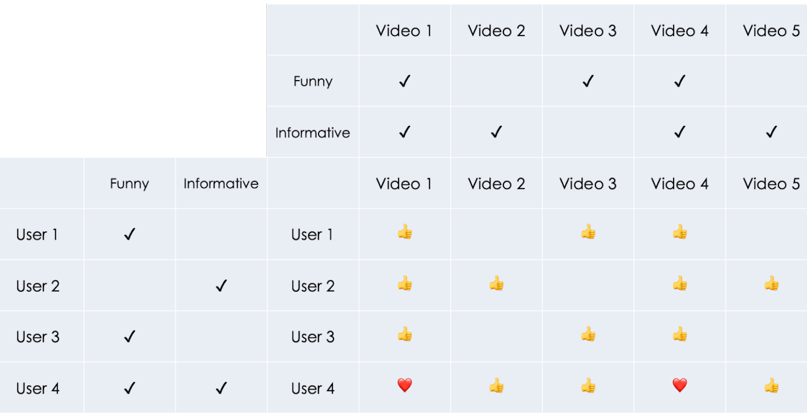
The first generation of large-scale recommendation algorithms, such as Amazon and Netflix in the early 2000s, used a simple technique called collaborative filtering: “Customers who bought this also bought that.”[45] To recommend items to a user on their homepage when they’re not looking at any particular item, simply take the recommendation lists of the items they’ve viewed or bought in the past, and aggregate the lists in some sensible way. Although crude by today’s standards, collaborative filtering proved powerful in e-commerce, sometimes finding surprising correlations between products. The story about supermarkets putting beer next to diapers to cater to frazzled fathers is apocryphal but accurately conveys the idea that the purchase data might reveal non-obvious connections.[46]

In 2006, Netflix publicly released movie ratings by half a million of its users, totaling 100 million ratings, and asked researchers to use this data to improve its recommendation algorithm. The most accurate algorithm would win a million-dollar prize. The contest supercharged the research field—not because of the prize, but because it was by far the largest such dataset available publicly.

The big insight to come out of the contest was the idea of matrix factorization, which undergirded what I see as the second generation of recommendation algorithms. Here’s the idea, illustrated with a hypothetical toy example. In this matrix describing the past engagement scores of a few users with a few videos, there are clearly many patterns. What might explain them?

	Video 1	Video 2	Video 3	Video 4	Video 5
User 1	👍		👍	👍	
User 2	👍	👍		👍	👍
User 3	👍		👍	👍	
User 4	❤️	👍	👍	❤️	👍

The following figure reveals it. To generate the data, I assumed that each video has two qualities: whether it’s funny and whether it’s informative. Some users like funny videos, some like informative videos, and some like both. If a video contains one attribute that a user likes, they give it a 👍. If it contains two, they give it a ❤️.



				Video 1	Video 2	Video 3	Video 4	Video 5
			Funny	✓		✓	✓	
			Informative	✓	✓		✓	✓
	Funny	Informative		Video 1	Video 2	Video 3	Video 4	Video 5
User 1	✓		User 1	👍		👍	👍	
User 2		✓	User 2	👍	👍		👍	👍
User 3	✓		User 3	👍		👍	👍	
User 4	✓	✓	User 4	❤️	👍	👍	❤️	👍

As oversimplified as this toy example seems, it turns out that real datasets show similar patterns. Of course, there are millions of users and posts, and hundreds or thousands of attributes (and user preferences regarding those attributes). And engagement can’t be exactly explained or predicted using those attributes: The predictions are merely correlated with the observed values, and there is a lot of noise. Most importantly, the matrices are “sparse”: Users have only ever come across a tiny fraction of the available posts, so most cells in the matrix are actually marked “N/A.”

Despite the size, noisiness, and sparsity of real-world datasets, it turns out that matrix factorization algorithms can identify—given only the matrix—a set of attributes and preferences which, when combined, result in approximately the same matrix. The algorithm can’t label those attributes with human-interpretable terms like “funny” or “informative,” but it doesn’t matter! Once it figures out those post attributes and user preferences, the algorithm can predict whether a user will like a post they’ve never encountered before.

This idea was a revolution in recommender systems research. Contestants who used it shot to the top of the Netflix Prize leaderboard, and its value became apparent. By the end, all the top contestants, including the winner, used it. Note that this algorithm uses only behavioral records and completely ignores user demographics and movie metadata such as genre. In fact, given large enough behavioral records, it will automatically discover genres as attributes underlying the matrix![47] The algorithm would refer to the attributes by opaque IDs rather than labels like “comedy” or “action,” but again, it doesn’t matter if the only goal is prediction.

Matrix factorization is my favorite example of research that wows people in the lab but doesn’t translate well to the real world. Unexplainable predictions were just fine for the contest but didn’t make for a great user experience. Being told you’ll like a movie without being told why is unsatisfying. Look at Netflix today: It labels recommendations with categories like “feel-good comedy dramas,” for good reason.

Besides, for social media, matrix factorization is a nonstarter. Netflix, at the time, had a tiny inventory of about 18,000 videos, so the algorithm was possible to compute. On a scale of billions of posts, it is computationally intractable, especially considering that the algorithm has to work in real time as new posts are constantly being uploaded. It’s possible that computational difficulty might be one reason why many social media platforms were late to the recommendation game. Due to the dominance of matrix factorization in the research world in the late 2000s, they may have rejected the entire approach as infeasible.

But once social media platforms started adopting recommendation algorithms, there was no turning back. By the time ByteDance launched TikTok in 2016, recommendation algorithms were successful enough that making the algorithm the core of the product would have been an obvious idea. Interestingly, ByteDance and its founder Zhang Yiming are reported to have launched a series of products going back to 2012 with the same concept: algorithmic content recommendations without a social network.[48]

Today there is a diversity of algorithms in use. One powerful technique is to “embed” users and posts as points in a high-dimensional space (with, say, a few hundred or a few thousand dimensions).[49] Distances between users and posts roughly capture the idea of similarities in attributes and preferences. Deep learning is usually, but not always, used to learn the embedding: the mapping from a behavioral record to a point in the high-dimensional space.[50]

It's worth pausing to ask how well recommendation algorithms work. It may seem obvious that they must work well, considering that they power tech platforms that are worth tens or hundreds of billions of dollars. But the numbers tell a different story. One way to quantify it is by engagement rate: the likelihood that a user engages with a post that is recommended to them. On most platforms, it is less than 1%. TikTok is an outlier, but even there, it is only a little over 5%.[51] This is not because the algorithms are bad, but because people are just not that predictable. As I've argued elsewhere, when the user interface is good enough, users don't mind the low accuracy.[52]

Does that mean that all this algorithm talk is nonsense? If they were so hit-or-miss, how can recommendation algorithms possibly be causing all that is attributed to them? Well, even though they are imprecise at the level of individual users, they are accurate in the aggregate. Compared to network-based platforms, algorithmic platforms seem to be more effective at identifying viral content (that will resonate with a large number of people). They are also good at identifying niche content and matching it to the subset of users who may be receptive to it. I believe it is in the aggregate sense that recommendation algorithms are most powerful—and sometimes dangerous.

WHAT THE ALGORITHM ISN'T

SOCIAL MEDIA COMPANIES have hired hundreds of psychologists.[53] So it's easy to imagine that platform algorithms have programmed into them an array of insights about how to recommend content that hooks us. That's not the case. Behavioral expertise is useful in designing the *user interfaces* of apps, but there is little human knowledge or intuition about what would make for a good recommendation that goes into the design of their *algorithms*. The algorithms are largely limited to looking for patterns in behavioral data. They don't have common sense.

This can lead to algorithmic absurdities: like ads featuring earwax, toenail fungus, or other disgusting imagery.[54] Again, it's easy to imagine that this is the result of (devious) intent: evil advertisers spending money just so they can cackle in the knowledge that millions of people around the world are getting grossed out.

But it is almost certainly the result of algorithmic optimization of click-through rates (which advertisers have learned to exploit for their own purposes). The key thing to remember is that the click through rate for ads is infinitesimal.[55] So if even, say, 0.1% of people click on gross-out ads for whatever reason—morbid curiosity?—the ad engines count that as success. They don't see and don't care about the people who hit the back button as soon as they see the image. This harms the publisher in addition to the user, but neither party has any much power to change things.

Although the approach of optimization based on machine learning is ubiquitous today, it wasn't always the case. Take Facebook. Back in 2010, it used an algorithm it called "EdgeRank" to construct a user's feed. This is what it looked like:[56]

```
priority(user,item)=affinity(user,poster)*Weight[item.type]/item.age
```

This formula is invoked for every item that can potentially be shown to the user, i.e., content posted or shared by their friends, events in the user's area, etc. Items are shown in decreasing order of priority, likely with a few additional tweaks not captured in this formula.

The two key ingredients in the formula are the affinity score and the item type weights. The affinity score represents Facebook's prediction of how much the user in question wants to see updates from the poster. This was again a manually programmed formula that took into account things like whether the user recently interacted with the poster; no machine learning was involved. The item type weight reflected Facebook engineers' predictions regarding the type of content that was more engaging: photos more than text, for example. These were also manually set rather than learned. To be clear, the manual judgments I refer to are about broad types of content, such as photos, comments, events, and so on. They are not at the level of individual users or posts, or even categories of users or posts such as "elected officials" or "breaking news."

The inverse dependence[57] of priority on the age of the item means that newer items are more likely to be at the top. But this is not a strict relationship: An older item from a poster with high affinity to the user can still end up on top. That means that the feed was roughly reverse chronological, but not exactly.

CASE STUDY: MEANINGFUL SOCIAL INTERACTIONS

EdgeRank didn't last long and was replaced by machine learning. In 2018, Facebook introduced a metric called "meaningful social interactions (MSIs)" to the machine learning system. The stated goal was to decrease the presence of media and brand content in favor of friends-and-family content.[58] Here is my best understanding of how the algorithm worked, pieced together from Facebook's own high-level descriptions and low-level details in the Haugen documents.[59]

$$\texttt{MSI(user, item)} = \texttt{affinity(user, poster)} * \sum_{\texttt{int-type}} \texttt{P(user, item, int-type)} * \texttt{Weight[int-type]}$$

The formula calculates a meaningful social interaction score for each item that could be shown to a given user. As before, the feed is created by ordering available posts roughly by decreasing MSI score, but with tweaks for things like diversity. P(user, item, int-type) is the predicted probability that the user will have a specific type of interaction (such as liking or commenting) with the given item. These probabilities are predicted using a machine-learned model. The probability calculation is the workhorse of the algorithm and is where the sophistication of the system resides (for example, if Facebook were to use matrix factorization, it would be to calculate these probabilities). The MSI formula computes a weighted sum of those probabilities; there is a manually defined set of weights for each interaction such as liking or commenting, which we'll discuss in a moment. Finally, the MSI formula adjusts the result based on the affinity of the user to the poster.

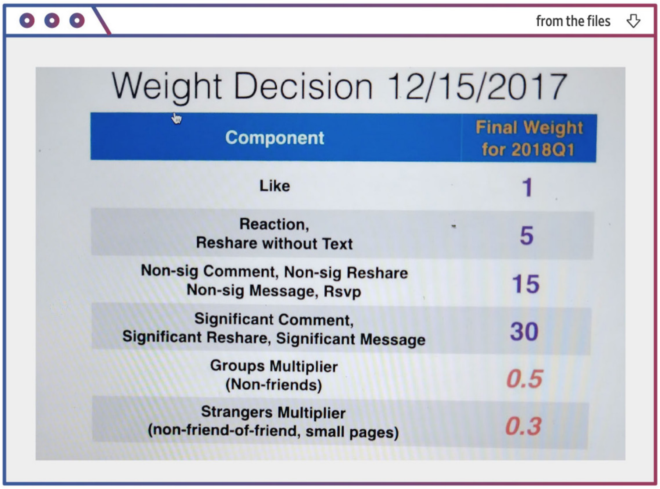
There are fewer knobs for engineers to tweak here than in EdgeRank, and the core logic—the engagement probability calculation—is handled via machine learning. There are no longer manual weights for item types like photos or videos. If some types of content are more engaging than others, that will be automatically learned from the data—in fact, it will be learned on a per-user level, so that the algorithm may prefer photos for one user and events for another.

Taking this logic to its natural conclusion, there should be no need to manually adjust the formula by affinity. If users like to see content from friends over brands, the algorithm should be able to learn that—again, at a granular, per-user level that cannot be achieved by manual tweaking of weights. Why, then, does the formula use affinity scores? It appears to be an explicit attempt to fight the logic of engagement optimization, manually programming in a preference for friends-and-family content even at the expense of short-term engagement with the aim of increasing long-term satisfaction, which the algorithm can't measure.

PLAYING WHAC-A-MOLE WITH HANDS TIED BEHIND THEIR BACKS

IT IS DEBATABLE how much control engineers have over the effects of recommendation algorithms. My view is that they have very little. Let me illustrate with an example. In 2019, Facebook realized that viral posts were much more likely to contain misinformation or many other types of harmful content.[60] (The correlation between virality and misinformation is also consistent with some research.)[61] In other words, the shift to Meaningful Social Interactions had had the opposite of the intended effect: Content that provoked outrage and stoked division was gaining in reach instead. This was a key point of Frances Haugen's testimony and has been extensively written about in the press.[62]

Figure 7: Interaction type weights and affinity multipliers for the Meaningful Social Interactions formula, via WSJ.[63]



Weight Decision 12/15/2017

Component	Final Weight for 2018Q1
Like	1
Reaction, Reshare without Text	5
Non-sig Comment, Non-sig Reshare Non-sig Message, Rsvp	15
Significant Comment, Significant Reshare, Significant Message	30
Groups Multiplier (Non-friends)	0.5
Strangers Multiplier (non-friend-of-friend, small pages)	0.3

Source: December 2017 internal Facebook memo titled 'The story of deriving the Meaningful Social Interactions metric weights (UX Research & Data Science)'

A look at the table of weights for the MSI formula instantly reveals a potential reason for this. Resharing a post was weighted 30 times higher than liking it. The logic behind such a high weight is presumably to identify posts that were potentially viral and boost them even further.

After recognizing the unintended consequence of this weight decision, in 2020, Facebook dropped the reshare weight all the way to 1.5. But the weight for *comments* remained high. Whereas reshares and comments were grouped in a single category in 2018, they no longer were. So here's how the weights looked in the first half of 2020. (There are few documents after this date in the Facebook files, and some minor changes are mentioned, but it is not clear whether and when there were any major weight updates after this date.)

Table 3: Interaction-type weights for the MSI formula in 2020.

Interaction type	weight
Like	1
Reaction	1.5
Reshare	1.5
Comment	15-20

Comments are overwhelmingly more important than any other type of interaction. Although it doesn't seem to have been reported in the press, a likely consequence of these weights is that posts that implicitly or explicitly encouraged users to comment would have done even better after this change. And one reliable way to encourage people to comment is to post divisive content. Fox News had long employed this strategy. According to one former social media producer: "We would intentionally post content that would be divisive and elicit a lot of comments."[64]

In short, it's quite possible that just as Facebook's attempt to clamp down on harmful content by moving to MSI backfired, its 2020 attempt to fix the problems with MSI *also* backfired. We can't know for sure unless there are more revelations of internal documents. It shouldn't be surprising, though, that attempting to steer a system of such extraordinary complexity using so few knobs would prove challenging.

I think there are two driving principles behind Facebook engineers' thinking that explain why they've left themselves with so little control. I'm basing this on the design of the algorithm, the internal discussions about it in the Haugen documents, and the generally prevalent views on these questions among technologists.

First, the system is intended to be neutral towards all content, except for policy violating or "borderline" content. To be sure, the list of types of content, users, and groups that are algorithmically demoted seems to be ever growing. The January 6 committee's draft report on social media lists dozens of such actions that Facebook took leading up to and following the 2020 U.S. election (which was apparently not enough!).[65] But no matter how much content is demoted, it is not the same as having a thesis about what types of content should thrive on the platform, and designing around such a thesis.[66]

There are obvious and important arguments in favor of neutrality. After all, platforms are already under attack from all sides of the political aisle for supposedly being biased. But neutrality is hard to achieve in practice. Many biases are emergent effects of these systems. One is the rich-get-richer effect: Those who already have a high reach, whether earned or not, are rewarded with more reach.[67] For example, the top 1% of authors on Twitter receive 80% of tweet views.[68] Another is demographic bias: Users' tendency to preferentially engage with some types of posters may be amplified by the algorithm.[69] Ultimately, designing for neutrality ends up rewarding those who are able to hack engagement or benefit from social biases.

The second main driving principle is that the algorithm knows best. This principle and the neutrality principle reinforce each other. Deferring the policy (about which content to amplify) to the data means that the engineers don't have to have a point of view about it. And that neutrality in turn provides the algorithm with cleaner data from which to learn.

The algorithm-knows-best principle means that the same optimization is applied to all types of speech: entertainment, educational information, health information, news, political speech, commercial speech, art, and more.[70] If users want more or less of some types of content, the thinking goes, the algorithm will deliver that. The same applies to any other way in which a human designer might try to tweak the system to make the user experience better. For example, suppose someone suggested to a Facebook engineer that posts related to the user's job or career, posts from colleagues, etc. should have slightly higher priority during work hours, with posts about parties or entertainment prioritized during evenings or weekends. The engineer might respond along these lines: "But who are we to make that decision? Maybe people actually secretly want to goof off during work. If so, the algorithm will let them do that. But if the policy you're asking for is actually what users want, then the algorithm will automatically discover it."

The level of adherence to these principles can be seen in how timid the deviations are. For example, in early 2021, Facebook trained machine learning models to classify posts as "good for the world" or "bad for the world."[71] The training data was generated by surveying users. Facebook found that posts with higher reach were more likely to be "bad for the world," so it wanted to algorithmically demote them. The first model that it built successfully suppressed objectionable content but led to a decrease in how often users opened the app—incidentally, an example of the ills of engagement optimization. So it deployed a tweaked, weaker model. What's most interesting to me about this is that, again, there's no articulated theory of what is good for the world. Even that judgment is delegated to the crowd. I mention this not to necessarily criticize it, but to point out that it is on one end of the spectrum of available governance approaches, and very different from traditional media. Even within the realm of democratic governance, there are other possible models that involve fewer people but more deliberation compared to crowdsourcing.[72]

HOW ENGAGEMENT OPTIMIZATION FAILS USERS, CREATORS, AND SOCIETY

MANY OF THE FAMILIAR pathologies of social media are, in my view, relatively direct consequences of engagement optimization. Understanding these connections will help us appreciate why reforms have proved difficult. The issues I identify in this section will persist even if companies improve transparency around their algorithms, invest more resources into content moderation, and provide users more control over what they see.

Let's start with how engagement optimization fails users. Behavioral data—data on past engagement—is the critical raw material for recommendation engines. The more data, the better the model is able to drive future engagement. So platforms emphasize feedback types that are more frequent. An example of this viewpoint from YouTube researchers in 2016: "Although explicit feedback mechanisms exist on YouTube (thumbs up/down, in-product surveys, etc.) we use the implicit feedback of watches to train the model, where a user completing a video is a positive example. This choice is based on the orders of magnitude more implicit user history available. …"[73] This is generally true across platforms, and over time, there has been a shift to “implicit” forms of feedback where the user action is minimal.[74] On TikTok, users don’t need to select videos, just swipe.

The problem with implicit feedback is that it relies on our unconscious, automatic, emotional reactions: “System 1,” rather than our rational and deliberative mode of thought: “System 2.”[75] A rich literature in behavioral economics documents the biases that System 1 suffers from. A TikTok user might swipe past a video by a medical expert reminding people to get a flu shot because she doesn’t look like the stereotype of a medical expert, and dwell on an angry video that they enjoy in the moment but regret later. By default, implicit-feedback-based feeds cater to our basest impulses.

From the perspective of creators, the most salient fact about engagement optimization is that it is a fickle overlord. If a creator puts out two pieces of content, one of which the data shows to be 10% more engaging than the other, the algorithm will reflect that in its recommendations and will compound that difference over time. The consequence might be that the first piece of content has a hundred times the reach of the other. The high variance and unpredictability of reach means that commercial content creators face an uncertain revenue stream on algorithmic platforms. They adapt in various ways. Producing a large volume of content, even if lower quality, can increase the chances that at least a few will go viral each month and smooth out the revenue stream.[76]

Still, an environment where everyone is looking for the next viral hit makes it hard for creators to be authentic. It leaves little room for those who are uninterested in popularity and simply want to have a conversation with a small group of people. The increase in distribution of viral content comes at the expense of suppressing more boutique types of content. Note that one appeal of non-algorithmic platforms such as Substack is the reliability and up-front predictability of revenue.[77]

Let’s turn to harm to society: specifically, harms that cannot be understood as the aggregate of harms to individuals. In other words, while it may be true that social media use harms mental health, I view that as a harm to individuals. The difference matters. The familiar complaints about social media, such as privacy and exploitation, aren’t particularly compelling if viewed as transactional harms to individuals, but far more serious from a collective, structural perspective.[78]

I want to highlight one particular set of harms to society, pertaining to institutions and markets: institutions like science, journalism, public health, and art, and markets like restaurants or travel. Each of these institutions and markets has been reshaped to varying degrees through algorithmic platforms. This is most obvious in the case of news, which is heavily reliant on platforms for distribution. It is starting to happen with science and scholarship, as social media has become a primary way for many of us to learn of our peers' work. While success on platforms might not affect whether a paper is accepted for publication, it does affect which papers will be heard about and built upon. As for the labor market, people often hear of job postings on social media.

Each institution has a set of values that make it what it is, such as fairness in journalism, accuracy in science, and aesthetic values in art. Markets have notions of quality, such as culinary excellence in restaurants and professional skill in a labor market. Over decades or centuries, they have built up internal processes that rank and sort what is produced, such as peer review. But social media algorithms are oblivious to these values and these signals of quality. They reward unrelated factors, based on a logic that makes sense for entertainment but not for any other domain.

As a result, I argue that social media platforms are weakening institutions by undermining their quality standards and making them less trustworthy. While this has been widely observed in the case of news,[79] my claim is that every other institution is being affected, even if not to the same degree. TikTok, best known for viral dances, might not seem like much of a threat to science. Maybe it won't be. But historically, we've observed that platforms start out as entertainment and gradually move into every sphere of speech.[80] That's already measurably true of TikTok for domains like politics: In a recent paper I co-authored, we counted over 2.5 million U.S. political videos by over 60,000 creators in a 45-day period preceding the 2020 election.[81]

THE LIMITS OF DATA SCIENCE

PLATFORM COMPANIES are well aware of these limitations. They've made occasional, rudimentary efforts to fix some of these issues—like Facebook's "bad for the world" classifier. Why haven't they done more? The most obvious explanation is that it hurts the bottom line. There's certainly some truth to this. The reliance on subconscious, automatic decision making is entirely intentional; it's called "frictionless design." The fact that users might sometimes exercise judgment and resist their impulses is treated as a problem to be solved.[82]

I don't think this is the entire answer, though. The consistent negative press has genuinely hurt platforms' reputation, and there have been internal efforts to do better. So it's worth talking about another limitation. Most of the drawbacks of engagement optimization are not visible in the dominant framework of platform design, which places outsize importance on finding a quantitative, causal relationship between changes to the algorithm and their effects. To explain what I mean, consider four reasons why someone might quit a social media app that they just took up.

1. The user consistently fails to get content recommendations that they found engaging enough. This is, of course, exactly what engagement optimization is designed to prevent.
2. The user got recommendations that were engaging in the moment but didn't make them feel good once they put down the app after a couple of hours. Companies are pretty good at optimizing against this outcome as well. A simple way to test an algorithm change is to A/B test it: that is, deploy it to a randomly selected subset of users. Then track how many of those users open the app on a given day, compared to a control group. Algorithms called contextual bandits automate some of the work of doing these A/B tests and tweaking the system based on their outcome.
3. The user enjoys the app, but over a period of six months, they realize that while it's a good source of entertainment, they haven't found it beneficial in any long-term way. This is tricky to measure! Of course,

platforms pay close attention to metrics like retention and churn, but countless changes are made over a period of six months, and without an A/B test, there's no good way to tell which changes were responsible for users quitting. And in most cases, you can't run an A/B test for six months: That's too slow. Still, for a few particularly important design decisions, platforms do employ long-running A/B tests. For example, Facebook found that showing more notifications increased engagement in the short term but had the opposite effect over a period of a year.[83]

4. The user's experience of the app as an individual is, on balance, positive at all time scales, but there has been a barrage of negative press about its harmful effects on other people and for democracy. The disconnect could be because individual users don't necessarily internalize societal harms: Users who consume election misinformation may actually love it. Or it could be because some harms such as privacy are structural and cannot be understood as the aggregate of individual, transactional harms.[84] At any rate, our hypothetical user quits because they decide that they don't want to help monetize the app given what they have heard about it in the press.

Measuring this is impossible even in theory! Experimenting on users critically relies on the assumption that each user's behavior is independent. Collective harms completely violate this assumption. Even if the platform were to run a yearslong A/B test, societal-scale harms such as undermining democracy affect *all* users (and nonusers), so the churn in the experimental group wouldn't necessarily be any higher than in the control group.

Table 4: Four levels at which platform algorithms may have effects, and ways in which those effects can potentially be measured. CTR = Click Through Rate. MSI = Meaningful Social Interactions, Facebook's engagement metric. DAU = Daily Active Users.

Level	Effect	Example metric	How to measure
Immediate	Engagement	CTR, MSI	Automated; optimized
Short term	User satisfaction	DAU	Manual (A/B test)
Long term	Value to user	Six-month retention	Too slow to measure
Global	Value to society	-	Impossible to measure

ALGORITHMS ARE NOT THE ENEMY

A **TEMPTING RESPONSE** to this litany of problems is to suggest that we should go back to chronological feeds. But this confuses the category of algorithmic recommendations with a specific kind of algorithm, namely engagement optimization. Of course, the only recommendation algorithms we've been exposed to are those that optimize for engagement, so it's hard to imagine any other kind. But to fail to do so would be to throw the baby out with the bathwater.

At their core, recommendation algorithms are a response to information overload: There is far more information online that is relevant to one's interests than one has time for. The problem is only getting worse. Chronological feeds were (barely) tenable a decade or two ago when user-generated online content was in its infancy. Today, offering only chronological feeds is not a realistic option for a mainstream platform that faces competitive pressures. Ranking algorithms are a practical necessity even in a purely subscription-based network, like Instagram a few years ago. The company has reported that by 2016, before it launched its algorithm, users missed 70% of all the posts in their feed.[85] Today, Instagram has five times as many users as it did then, so the overload problem would likely be even worse. Far from a return to chronological feeds, platforms are facing enormous commercial pressures to make algorithms even more central to their operation.

Search offers a useful analogy: Before search engines, people accessed online information through directories. I suspect that social media without recommendations will soon seem just as quaint if it doesn't already.

Let's also pause to consider the many benefits that algorithmic platforms have brought. The ability to go viral has enabled many creators, such as musicians and entertainers, to establish an initial livelihood on social media. This weakening of the power of gatekeepers has unleashed a creative energy that deserves to be celebrated.

Algorithmic recommendations excel at giving people niche content that they are interested in. Suppose I'm interested in learning about new restaurants in Princeton, New Jersey, where I live. What are my options? If I lived in a big city like New York City, there are many New York City foodie Instagram accounts I could follow. But Princeton is too small a market for maintaining a town foodie account to be worth anyone's time. This is no problem for TikTok. Knowing that I enjoy content about Princeton and content about food is enough for it to recommend content about Princeton restaurants from various accounts, even if each of them mostly posts content I'm not interested in (like food in central New Jersey or activities in Princeton).

Finally, let's keep in mind that "reverse chronological" is an algorithm, albeit a simple one. Chronological feeds are not normatively neutral: They are also subject to rich-get-richer effects, demographic biases, and the unpredictability of virality. There is, unfortunately, no neutral way to design social media. Algorithmic recommendations could in fact be an opportunity to actively counteract harmful patterns of information propagation.

CONCLUDING THOUGHTS

FOR ALL THE INK that's been spilled about social media algorithms, their role is only getting bigger. They're displacing social networking as the predominant method of amplifying speech. At the same time, they're displacing traditional forms of content moderation as the predominant method of suppressing speech. People interact with social media algorithms for several hours a day on average.[86] Beyond social media, recommendation algorithms have made their way into education (Coursera), finance (Robinhood), and many other domain-specific apps.

Yet recommendation algorithms remain poorly understood by the public. This knowledge gap has consequences ranging from mythologizing algorithms to policy stumbles.[87] Of course, algorithms aren't the whole picture: Just as important is the design of social media, platform processes, their incentive structures and, most critically, human-algorithm interactions. Demanding much more transparency from platform companies—and not being easily swayed by their arguments about competitive risks—will go a long way toward improving our understanding of all these aspects of social media.

Let's imagine a future where children learn how platform algorithms work, just as they learn about other types of civic infrastructure and grow up empowered to participate in a healthier way on algorithmic platforms, as well as to help govern them.

NOTES

[1] Dell Cameron et al., *Facebook Papers Directory*, Gizmodo (Feb. 14, 2022), https://gizmodo.com/facebook-papers-how-to-read-1848702919 [https://perma.cc/6PL2-X8QQ].

[2] Donella H. Meadows, Thinking in Systems: A Primer (Chelsea Green Publishing, 2008).

[3] *Braess's Paradox*, Wikipedia (Dec. 29, 2022, 12:30 PM), https://en.wikipedia.org/wiki/Braess%27s_paradox [https://perma.cc/X5HT-SDQY].

[4] Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest (Yale University Press, 2017).

[5] Jeremy A Frimer et al., *Incivility Is Rising Among American Politicians on Twitter*, 14 Soc. Psych. & Personality Sci. (2023).

[6] Nathan Matias & Lucas Wright, Impact Assessment of Human-Algorithm Feedback Loops (Just Tech. Social Science Research Council, 2022), https://just-tech.ssrc.org/field-reviews/impact-assessment-of-human-algorithm-feedback-loops/ [https://perma.cc/65Z2-AVUX].

[7] While friending/following is about more than content propagation, I've put friend recommendation under that group to keep the categorization simple.

[8] *Never Miss Important Tweets from People You Follow*, Twitter (Feb. 10, 2016), https://blog.twitter.com/official/en_us/a/2016/never-miss-important-tweets-from-people-you-follow.html [https://perma.cc/WG3J-8P8M].

[9] Alex Hern, *How TikTok's Algorithm Made It a Success: 'It Pushes the Boundaries,'* The Guardian (Oct. 24, 2022, 1:00 AM), https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity [https://perma.cc/NQ7C-9JPR].

[10] On mobile devices, the Discover feed appears below the search box on Google.com, on the Chrome new tab page, and on the home screen on Android. It is not accessible on desktop devices. Google doesn't seem to use the name Discover in the product itself and simply calls it "Google," which is perhaps one reason why so little has been written about it.

[11] Joan E. Solsman, *YouTube's AI Is the Puppet Master Over Most of What You Watch*, CNET (Jan. 10, 2018, 10:05 AM), https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/ [https://perma.cc/R5LD-JMNJ].

[12] Kalley Huang & Mike Isaac, *Instagram Rolls Back Some Product Changes After User Backlash.*, N.Y. Times (July 28, 2022), https://www.nytimes.com/2022/07/28/technology/instagram-reverses-changes.html.

[13] Adam Mosseri, *Shedding More Light on How Instagram Works*, Instagram (June 8, 2021), https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works [https://perma.cc/ZR9V-JW82].

[14] Jameel Jaffer (@JameelJaffer), Twitter (Sept. 19, 2022, 9:07 AM), https://twitter.com/jameeljaffer/status/1571848466174713856; Joe Biden (@JoeBiden), Twitter (Sept. 26, 2022, 4:15 PM), https://twitter.com/joebiden/status/1574492776988917764?lang=en.

[15] Sharad Goel et al., *The Structural Virality of Online Diffusion*, 62 Mgmt. Sci. 180, 181 (2016).

[16] *Id.* at 182-83.

[17] Travis Martin et al., Exploring Limits to Prediction in Complex Social Systems, Cornell Univ., ArXiv (2016), https://arxiv.org/abs/1602.01013 [https://perma.cc/2N44-L45W].

[18] Benjamin Guinaudeau et al., *Fifteen Seconds of Fame: TikTok and the Supply Side of Social Video*, 4 Computational Commc'n Rsch. 463 (2022).

[19] The authors of the structural virality paper utilized access to the Twitter "Firehose" API, which is no longer available.

[20] More visualizations of different accounts from different platforms are available at https://www.cs.princeton.edu/~arvindn/distorted-speech/bubble_chart/ [https://perma.cc/UJ7U-MPMU].

[21] Tarleton Gillespie, *Do Not Recommend? Reduction as a Form of Content Moderation*, 8 Soc. Media + Soc'y (2022).

[22] Not all platforms use the term engagement to describe what they optimize for. But I think it is fair to apply that term, as it is commonly (and flexibly)

understood, to most major platforms.

[23] Julia Carrie Wong, *Facebook Overhauls News Feed in Favor of 'Meaningful Social Interactions,'* The Guardian (Jan. 11, 2018. 9:31 PM), https://www.theguardian.com/technology/2018/jan/11/facebook-news-feed-algorithm-overhaul-mark-zuckerberg [https://perma.cc/SED9-3NUV].

[24] Smitha Milli et al., *From Optimizing Engagement to Measuring Value*, Cornell Univ., ArXiv (2021), https://arxiv.org/abs/2008.12623 [https://perma.cc/YRH3-NTNR].

[25] Paul Covington et al., *Deep Neural Networks for YouTube Recommendations* (2016), https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf [https://perma.cc/CU2L-4Z4K].

[26] Eric Meyerson, *YouTube Now: Why We Focus on Watch Time*, YouTube Off. Blog (Aug. 10, 2012), https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/ [https://perma.cc/82HQ-7WP6].

[27] Ben Smith, *How TikTok Reads Your Mind*, N.Y. Times, (Dec. 5, 2021), https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html.

[28] *An Update on Our Work to Safeguard and Diversify Recommendations*, TikTok Newsroom (Dec. 16, 2021), https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-safeguard-and-diversify-recommendations [https://perma.cc/S96X-HY34].

[29] Guinaudeau et al., *supra* note 18.

[30] *Netflix Prize*, Wikipedia (Feb. 21, 2023, 3:55 PM), https://en.wikipedia.org/wiki/Netflix_Prize [https://perma.cc/59LX-6NLQ].

[31] Xavier Amatriain & Justin Basilico, *Netflix Recommendations: Beyond the 5 Stars (Part 1)*, Netflix Tech. Blog (Apr. 6, 2012), https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429 [https://perma.cc/NZ4F-FUGY].

[32] Alex Heath, *Facebook is Changing Its Algorithm To Take on TikTok, Leaked Memo Reveals*, The Verge (June, 15, 2022, 12:46 PM), https://www.theverge.com/2022/6/15/23168887/facebook-discovery-engine-redesign-tiktok [https://perma.cc/X9GF-8Y6Q].

[33] Luke Thorburn et al., *How Platform Recommenders Work*, Medium (Jan. 20, 2022), https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a [https://perma.cc/3WCQ-GCWK].

[34] Amatriain & Basilico, *supra* note 31; James McInerney et al., *Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits*, ACM Digital Library (2018), https://static1.squarespace.com/static/5ae0d0b48ab7227d232c2bea/t/5ba849e3c83025fa56814f45/1537755637453/BartRecSys.pdf [https://perma.cc/GWC9-XRH8].

[35] Sahin Cem Geyik et al., *Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search*, Cornell Univ., ArXiv (2019), https://arxiv.org/abs/1905.01989 [https://perma.cc/VY7U-2D4D].

[36] Arvind Narayanan, *TikTok's Secret Sauce*, Knight First Amend. Inst. at Colum. Univ. (Dec. 15, 2022), https://knightcolumbia.org/blog/tiktoks-secret-sauce [https://perma.cc/5UYS-7HW5].

[37] Dmitry Pashtukov, *Inside Spotify's Recommender System: A Complete Guide to Spotify Recommendation Algorithms*, Music Tomorrow Blog (Feb. 9, 2022), https://www.music-tomorrow.com/blog/how-spotify-recommendation-system-works-a-complete-guide-2022 [https://perma.cc/4BNE-F7RG].

[38] Narayanan, *supra* note 36.

[39] Eytan Bakshy et al., *Exposure to Ideologically Diverse News and Opinion on Facebook*, 348 Sci. (2015); Ferenc Huszár et al., *Algorithmic Amplification of Politics on Twitter*, 119 Proc. of Nat'l Acad. of Sci. (2022).

[40] Venu Satuluri et al., SimClusters: *Community-Based Representations for Heterogeneous Recommendations at Twitter*, 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (2020), https://dl.acm.org/doi/proceedings/10.1145/3394486 [https://perma.cc/3RFK-RTXK].

[41] Tanya Dua, *Never-Before-Seen TikTok Stats from Leaked Presentations Show How It's Trying To Lure Adertisers to the Platform*, Bus. Insider (2021), https://www.businessinsider.com/tiktok-pitch-deck-shows-new-e-commerce-ads-2021-4 [https://perma.cc/L6RZ-45MS].

[42] Lara Zarum, *Some Viewers Think Netflix Is*

Targeting Them by Race. Here's What to Know, N.Y. Times (Oct. 23, 2018), https://www.nytimes.com/2018/10/23/arts/television/netflix-race-targeting-personalization.html.

[43] Traditionally, content and behavior corresponded to two types of recommendation algorithms called content filtering and collaborative filtering, but there is no reason to use only one type of signal, so the distinction has gradually become meaningless.

[44] Clive Thompson, *If You Liked This, You're Sure To Love That*, N.Y. Times (Nov. 21, 2008), https://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html.

[45] *Id.*; Greg Linden et al., *Amazon.com Recommendations: Item-to-Item Collaborative Filtering*, 7 IEEE Internet Computing (2003).

[46] For actual examples of surprising correlations in Netflix's data, see Libby Plummer, *This Is How Netflix's Top-Secret Recommendation System Works*, Wired UK (Aug. 22, 2017, 7:00 AM), https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like [https://perma.cc/8P3Z-DBF8].

[47] Robert M Bell et al., *All Together Now: A Perspective on the Netflix Prize*, 23 Chance 24 (2010), https://chance.amstat.org/2010/02/netflix/ [https://perma.cc/BH4N-VGZV].

[48] Alex W. Palmer, *How TikTok Became a Diplomatic Crisis*, N.Y. Times (Dec. 20, 2022), https://www.nytimes.com/2022/12/20/magazine/tiktok-us-china-diplomacy.html.

[49] Confusingly, in the computer science literature, embedding spaces are referred to as low-dimensional because this is with reference to the number of possible dimensions, which is equal to the number of users or posts.

[50] Sally Goldman, *Embeddings*, Google Devs. (2022), https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture [https://perma.cc/S67M-J84U].

[51] Elena Cucu, *[STUDY] TikTok Benchmarks: Performance Data and Stats Based on the Analysis of 616,409 TikTok Videos*, Socialinsider Blog (Sept. 21, 2022), https://www.socialinsider.io/blog/tiktok-benchmarks/ [https://perma.cc/86T4-29T2].

[52] Narayanan, *supra* note 36.

[53] Alvaro M. Bedoya, Fed. Trade Comm'n, Address at the National Academy of Sciences, Engineering & Medicine Meeting of the Committee on the Impact of Social Media on the Health and Wellbeing of Children & Adolescents (Feb. 7, 2023), https://www.ftc.gov/system/files/ftc_gov/pdf/national-academies-speech-bedoya.pdf [https://perma.cc/42GH-95YY].

[54] Shane Hegarty, *In Using Ad Blockers and the Default Age Setting, Google Thinks I'm 120+ with an Earwax Problem and Some Dodgy Interests*, Times (London) (Oct. 17, 2021), https://www.thetimes.co.uk/article/ad-blockers-default-age-setting-google-ads-earwax-problem-r9howq5md; Sarah Cooper (@sarahcpr), Twitter (Oct. 25, 2020, 2:03 PM), https://twitter.com/sarahcpr/status/1320425679410941953; Dave Ross, *When Will Google Let Us Block Those Repulsive Toenail Fungus Ads?* MyNorthwest.com (May 4, 2022, 6:26 AM), https://mynorthwest.com/3453078/ross-google-block-repulsive-toenail-fungus-ads/ [https://perma.cc/8GTF-WSC6].

[55] *What Is a Successful Click-Through Rate for Ads?* Broadstreet Ads (June 23, 2022), https://broadstreetads.com/successful-click-through-rate/ [https://perma.cc/3WJU-NW3M].

[56] This is based on Facebook's F8 developer conference in 2010, as documented by a third party. Jeff Widman, EdgeRank (2010), http://edgerank.net/ [https://perma.cc/CY4N-9NJV].

[57] It isn't clear if the inverse age dependence was linear or exponential, but the point stands either way.

[58] Mark Zuckerberg, Facebook (2018), https://www.facebook.com/zuck/posts/10104413015393571.

[59] Cameron et al., *supra* note 1.

[60] Max Reshare Depth Experiment (redacted for Congress), Documentcloud.org (2019), https://www.documentcloud.org/documents/21602015-tier1_rank_pr_1119.

[61] Soroush Vosoughi et al., *The Spread of True and False News Online*, 359 Sci. 1146 (2018); Michela Del Vicario et al., *The Spreading of Misinformation Online*, 113 Proc. of Nat'l Acad. of Scis. 554 (2016).

[62] Keach Hagey & Jeff Horwitz, *Facebook Tried*

to Make Its Platform a Healthier Place. It Got Angrier Instead., Wall St. J. (Sept. 15, 2021, 9:26 AM), https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215 [https://perma.cc/FE65-59YA]; *Facebook Whistleblower Testifies on Protecting Children Online*, C-SPAN (Oct. 5, 2021), https://www.c-span.org/video/?515042-1/whistleblower-frances-haugen-calls-congress-regulate-facebook.

63 Hagey & Horwitz, *supra* note 62.

64 David Uberti, *How Fox News Dominates Facebook in the Trump Era*, VICE News (Apr. 29, 2019, 10:20 AM), https://www.vice.com/en/article/wjvdem/how-fox-news-dominates-facebook-in-the-trump-era [https://perma.cc/8VKL-7FHT].

65 House Select Comm. to Investigate the Jan. 6th Attack on the U.S. Capitol, *Social Media & the January 6th Attack on the U.S. Capitol: Summary of Investigative Findings* (Draft, 2023), https://www.washingtonpost.com/documents/5bfed332-d350-47c0-8562-0137a4435c68.pdf [https://perma.cc/UG4Q-F42W].

66 For a detailed normative analysis of the difference between these approaches, see Seth Lazar, *Lecture 2. Communicative Justice and the Distribution of Attention* (Obert C. Tanner Lecture on Artificial Intelligence and Human Values, Jan. 18, 2023), https://write.as/sethlazar/cjda [https://perma.cc/BF6Q-HJML].

67 Linhong Zhu & Kristina Lerman, *Attention Inequality in Social Media*, Cornell Univ., ArXiv (Jan. 26, 2016), https://arxiv.org/abs/1601.07200 [https://perma.cc/72X9-M937].

68 Tomo Lazovich et al., *Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics*, 3 Patterns (2022), https://www.cell.com/action/doSearch?text1=Measuring+Disparate+Outcomes+of+Content+Recommendation+Algorithms+with+Distributional+Inequality+Metrics&field1=AllField&journalCode=patter&SeriesKey=patter [https://perma.cc/99S3-YHXU].

69 Christine Bauer & Andrés Ferraro, *Music Recommendation Algorithms Are Unfair to Female Artists, but We Can Change That*, The Conversation (Mar. 30, 2021, 8:58 AM), https://theconversation.com/music-recommendation-algorithms-are-unfair-to-female-artists-but-we-can-change-that-158016 [https://perma.cc/7R56-R3SP].

70 There are a few small exceptions. For example, after the 2020 U.S. elections, Facebook began experimenting with decreasing the amount of political content in news feeds (Anna Stepanov, *Reducing Political Content in News Feed*, Meta Newsroom (Feb. 10, 2021), https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ [https://perma.cc/AC3G-DP2U]). Later in 2021, the company made a more drastic attempt to demote all political content, but this had the unanticipated effect of suppressing high-quality news sources more than low-quality ones, and misinformation in fact rose (Jeff Horwitz et al., *Facebook Wanted Out of Politics. It Was Messier Than Anyone Expected.*, Wall St. J. (Jan. 5, 2023, 9:51 AM), https://www.wsj.com/articles/facebook-politics-controls-zuckerberg-meta-11672929976 [https://perma.cc/R5MP-A2L2]).

71 Kevin Roose et al., *Facebook Struggles to Balance Civility and Growth*, N.Y. Times (Jan. 7, 2021), https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

72 Aviv Ovadya, *Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure*, Harv. Kennedy Sch., Belfer Ctr. for Sci. & Int'l Affs. (Oct. 18, 2021) https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure [https://perma.cc/46QE-ETMB].

73 Covington et al., *supra* note 25.

74 Ben Thompson, *Instagram, TikTok, and the Three Trends*, Stratechery (Aug. 16, 2022), https://stratechery.com/2022/instagram-tiktok-and-the-three-trends/ [https://perma.cc/8Q2E-WRGS].

75 Daniel Kahneman, Thinking, Fast and Slow (Macmillan, 2011).

76 For example, one TikTok creator says, "In my experience you have to post 4-10 videos daily for one to pop off and get some traction on TikTok." Reddit user Ded___Pixel, Reddit (2022), https://www.reddit.com/r/ColinAndSamir/comments/vwf9hf/comment/ifpkhqa/ [https://perma.cc/6QXL-8UC4].

77 Substack - A New Model for Publishing, https://substack.com/going-paid [https://perma.cc/F6VE-

8NJ7] (last visited Mar. 7, 2023).

[78] Claire Benn & Seth Lazar, *What's Wrong with Automated Influence*, 52 CAN. J. OF PHIL. 125 (2022).

[79] Robyn Caplan & Danah Boyd, *Isomorphism Through Algorithms: Institutional Dependencies in the Case of Facebook*, 5 BIG DATA & SOC'Y (2018).

[80] Anders Olof Larsson, *The Rise of Instagram as a Tool for Political Communication: A Longitudinal Study of European Political Parties and Their Followers*, NEW MEDIA & SOC'Y (2021), https://journals.sagepub.com/doi/epub/10.1177/14614448211034158; Sam Bestvater et al., *Politics on Twitter: One-Third of Tweets from U.S. Adults Are Political*, PEW RSCH. CTR. (June 16, 2022), https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/ [https://perma.cc/8NVP-BPT8].

[81] Orestis Papakyriakopoulos et al., *How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok*, CORNELL UNIV., ARXIV (July 13, 2022), https://arxiv.org/abs/2206.04720 [https://perma.cc/LXM5-6QK4].

[82] Sam Lessin (@lessin), TWITTER (Aug. 9, 2022, 8:49 AM), https://twitter.com/lessin/status/1556986127785115648.

[83] Analytics at Meta, *Notifications: Why Less Is More—How Facebook Has Geen Increasing Both User Satisfaction and App Usage by Sending Only a Few Notifications*, MEDIUM (Dec. 19, 2022), https://medium.com/@AnalyticsAtMeta/notifications-why-less-is-more-how-facebook-has-been-increasing-both-user-satisfaction-and-app-9463f7325e7d [https://perma.cc/4E6A-ZZDM].

[84] Benn & Lazar, *supra* note 78.

[85] Mosseri, *supra* note 13.

[86] *Daily Time Spent on Social Networking by Internet Users Worldwide from 2012 to 2022*, STATISTA, https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ [https://perma.cc/U2FH-F9P4] (last visited Mar. 7, 2023).

[87] Kelley Cotter et al., *In FYP We Trust: The Divine Force of Algorithmic Conspirituality*, 16 INT'L J. OF COMM'N 2911 (2022); Daphne Keller, *Amplification and Its Discontents: Why Regulating the Reach of Online Content is Hard*, 21-05 KNIGHT FIRST AMEND. INST. AT COLUM. UNIV. (June 8, 2021), https://knightcolumbia.org/content/amplification-and-its-discontents [https://perma.cc/23KP-27GT].

About the Author

Arvind Narayanan is a professor of computer science at Princeton University, the incoming director of Princeton's Center for Information Technology Policy, and the 2022-2023 visiting senior research scientist at the Knight First Amendment Institute. He also led the Princeton Web Transparency and Accountability Project, which focused on uncovering how companies collect and use our personal information. Narayanan co-created a massive open online course and textbook on bitcoin and cryptocurrency technologies, which has been used in over 150 courses worldwide. His recent work has shown how machine learning reflects cultural stereotypes, and his doctoral research showed the fundamental limits of de-identification. Narayanan is a recipient of the Presidential Early Career Award for Scientists and Engineers, twice recipient of the Privacy Enhancing Technologies Award, and thrice recipient of the Privacy Papers for Policy Makers Award.

Acknowledgments

I'm grateful to Eric Chen, Alex Engler, Katy Glenn Bass, Jameel Jaffer, Sayash Kapoor, Soham Mehta, Roy Rinberg, and Deb Roy for feedback on a draft.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Emilie Flamme

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY