Testimony before the Senate Subcommittee on
Communications, Media, and Broadband

Hearing of December 9th, 2021

Algorithmic transparency and assessing effects of algorithmic ranking

Dean Eckles

Contemporary communication technologies have dramatically lowered monetary and practical costs of broadcasting information and media to many people — and of consuming others' broadcasts. They have created new ways for people to share their own thoughts, experiences, and creations, to consume and react to those shared by others, and — quite importantly — to rapidly propagate them. It is so easy to propagate content that people often share information that, on further consideration, they themselves would realize is misinformation.[1] Interactions on social media affect commerce, culture, politics, and public health,[2–6] thereby reasonably attracting scientific, public, and regulatory attention.

What role do algorithms play in all of this? Algorithms are unavoidable here. Even sorting posts by friends in chronological order* or videos by overall popularity is algorithmic; and often it is unclear there is a single, simple baseline algorithm. What much of the public conversation about algorithms has in mind are particular kinds of more complex, potentially more opaque, algorithms — typically based on statistical machine learning — that are adaptive to, e.g., features of the content and each person's history of consumption. In the context of social media, these algorithms typically first present items that are predicted to be something the consumer will take desirable actions on or would say is important, interesting, or fun.

Are these algorithms better or worse — for individual consumers and for society — than simpler alternatives that would, e.g., present everything from the accounts someone follows in chronological order? Can we straightforwardly specify what a given algorithm "amplifies" in a way suitable for assigning moral or legal responsibility?

Here I briefly summarize some key points.

1. At established platforms, algorithmic ranking and recommendation involve using many signals and are typically not aimed at simply maximizing short-run engagement.

2. Quantifying the impacts of algorithmic ranking is quite difficult, even with access to proprietary data. This is not only because of the complexity of these technical systems, but due to people's complex and often strategic responses to changes in algorithms.

3. We lack clear evidence about broader benefits or harms of algorithmic ranking. Nonetheless, simple rankings and recommendations (e.g., chronological, overall popularity) can make some forms of undesirable strategic behavior easier.

4. Policy-makers can protect the ability of external researchers to probe these systems, and they can provide clear paths for platforms to retain and share data in privacy-preserving ways.

*For simplicity I use "chronological" to refer to a ranking that shows items ordered by recency — perhaps more precisely called reverse chronological.

The rest of this statement is organized as follows. First, I characterize current practice in algorithmic ranking and recommendation of content in social media. Second, I elaborate on how we can learn — and what we already know — about effects of algorithms in social media. Third, I conclude by stating some policy implications.

Current practice

Before considering assessing their effects or crafting policy, it can be useful to understand the state-of-the-art in algorithmic ranking and recommendation in social media. This involves choosing from a large collection of items (i.e. content, stories, posts, activity) that a user is eligible to see and determining which of those items are displayed in what order — and also how each item is displayed, as there are often multiple variations available. I start with a prototypical case, presenting a somewhat simplified solution, and then discuss a few relevant variations.

Consider the problem of choosing which of a large set of photos or videos shared by accounts a user follows to display and in what order. This is a version of the problem faced by Instagram and Snapchat in their feeds (that is, setting aside their other channels for now). People do not typically spend enough time using these services to see all — or even the majority — of what is shared by the accounts they follow.[7,8]

These platforms have multiple goals in mind when doing this ranking. They then attempt to quantify these goals in metrics, usually defined at the level of each user. These could include the fraction of days that users log in, the number of photos they post, their time spent using the service, what they would say in response to a survey question asking about the service, the revenue from them viewing or clicking on ads, etc. These may be combined into a single evaluation criterion,[9] or managers may decide to try to maximize one metric while ensuring that they do not have substantial negative effects on others.

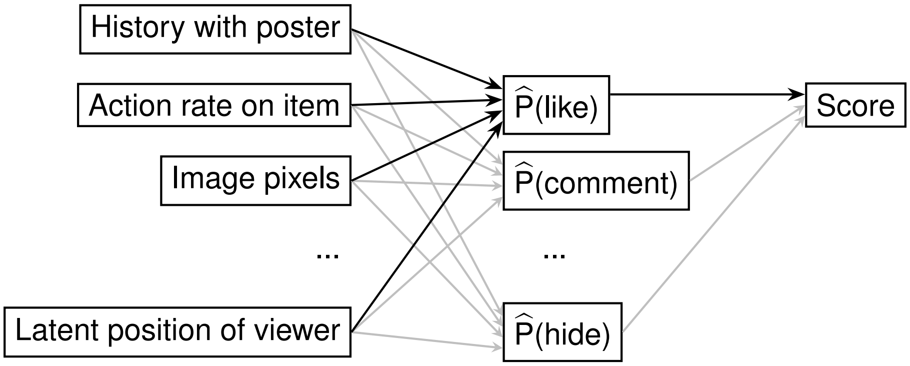
It is difficult to directly optimize what items to show to maximize such metrics, since these metrics are defined at the level of users and there are many possible rankings of items.[*] So typically platforms simplify the problem by defining a small number of constituent scores for each item, defining a combined score that, e.g., sums up these scores, and then ranking items by that combined score (Figure 1).[10,11]

These constituent scores are typically predictions of actions the viewing user might take on or related to the item. For example, there could be predictions of the probability the viewer will "like" the item.[†] Similar predictions can be made for many other item-level actions, such as commenting on it or viewing it for at least *x* seconds. Usually

[*] For a single user, there often are more possible rankings of available items than there are atoms in the universe: say there are 100 items available, then there are over 10^{156} possible orderings of those items.

[†] Ranking engineers at multiple firms can be overheard referring to "p-like".[12]

Figure 1: Schematic representation of inputs to scores used to rank items in social media feed. Numerous characteristics of the item are used to predict the probability of the viewer taking various actions (e.g., "liking" the item) based on historical data. These predictions together determine a combined score; for example, the combined score might add up these predictions, each with a different weight.



"negative" actions, such as unfollowing or unfriending the poster or hiding the item, are similarly predicted and given negative weights in the combined score.* Data scientists and managers try to identify new actions to predict that provide additional relevant signals (e.g., sequences of actions indicative of "regretted" clicks).

The same approach is often used to predict other, less direct, actions. For example, the platform may produce predictions of whether the viewer will follow or friend someone involved in the item (e.g., someone tagged in the photo). The general approach can also incorporate non-behavioral signals that are only available for a small sample of users, such as data from surveys. For example, if the platform has contractors or everyday users rating content in their feeds,† they can predict those ratings; that is, the platform can produce predictions of whether, if asked, the viewer would say this item is important, informative, funny, or makes them feel connected.

How are the weights on these different constituent scores chosen? In some cases these weights may be selected through managerial judgement alone, but often sophisticated A/B tests‡ are used to compare how different choices of weights affect metrics.[10,14,15] For example, the weights might be selected to maximize a metric designed to (ambitiously) measure people's level of "meaningful social interaction", subject to a constraint on revenue. While it is most straightforward to assess effects on metrics for the viewing user, many of the immediate

*That is, the combined score for an item will look something like score = $w_{\text{like}} \times \widehat{P}(\text{like}) + w_{\text{comment}} \times \widehat{P}(\text{comment}) + \ldots + w_{\text{hide}} \times \widehat{P}(\text{hide})$, where the ws are the (positive or negative) weights given to the predictions of different actions. Recent reporting makes clear TikTok's ranking likewise follows this pattern,[12] though it apparently lacks both any negative signals or any signals from attitudinal (e.g., survey) data.

†Facebook, for example, created a paid "feed quality panel" to meticulously rate items in their feeds, and has combined this with data from users who respond to prompts to rate their feeds.[13] These ratings can be used both to evaluate a prospective ranking algorithm and in the algorithm itself via predictions, as described here.

consequences of ranking apply to other people (e.g., by showing this item to this viewer, the poster of that item may receive an additional comment).[16] Platforms often try to incorporate these consequences into their ranking as well.

Recall the constituent scores are the probabilities of various actions; how do platforms predict these actions? They use very large statistical machine learning (or, if one prefers, artificial intelligence) models trained on historical data about which items people were shown, what the characteristics of those items where, and what the viewing user did.[*] For example, the historical data would generally reveal that if the viewer frequently comments on items involving a particular person (e.g., as the poster of the photo, as someone tagged in the photo), then they are more likely to do so for this new item. More recently, platforms have more thoroughly incorporated techniques from natural language processing and computer vision into these predictive models; the actual contents on the photo are used to predict what actions the viewer would take if shown it.[†] Furthermore, the history of a user's interactions and connections in the social network is sometimes used to learn some numeric representation of their preferences and dispositions, which can also be used in these predictions; these can be regarded as latent (i.e. unobserved) positions of each user.

Thus, the algorithmic ranking of social media feeds typically depends on numerous inputs. Some of these inputs themselves are the results of prior statistical machine learning (e.g., learned representations of the objects in photos). However, decision-makers within platforms are typically less focused on these inputs, but rather on understanding and effectively making tradeoffs between various metrics — defined not solely at the level of individual items to be ranked, but for individual users or for the entire service. They often make these tradeoffs by differently weighting predictions about how a viewer will act on or evaluate an item.

Further variations and complications

There are some variations on and complications of the above problem and solution.

1. When deciding what item to show in the second position, it can be useful to account for what the first item is; more generally, it can make sense to consider the full ranked set of items so as to reflect, e.g., demand for variety of topics or sources. While it is typically computationally difficult to directly optimize rankings in this way, platforms often implement various heuristics to improve the final ranking. For example, they may have rules that prevent a ranking

[‡] A/B tests (i.e., randomized experiments, randomized controlled trials) involve randomly assigning users to different variations on a service. Here this would be assigning users to rankings using different weights on the predicted actions. As in medicine and economics, these randomized trials are considered the "gold standard" for evidence and decision-making.

[*] Usefully, this historical data typically involves users seeing stories in an order other than the status quo because they are in an A/B test or because the scores for items have had a small amount of random noise added to them.

[†] That Facebook uses such signals can be seen in that ad delivery is immediately imbalanced by gender depending on whether the ad image alone includes stereotypically male- or female-relevant themes.[17]

from being highly repetitive in having several items from the same poster, only videos, etc.*

2. There are sometimes multiple ways to display an item. For example, multiple photos posted by the same person could all be shown smaller and grouped together, or some or all could be displayed at a larger size and shown individually. Some or all existing comments on an item could be displayed by default.[16] Thus, similar algorithms are also often used to make these decisions.

3. What comprises the inventory of items to be ranked is not always obvious, and it is often not limited to other accounts broadcasting new content (e.g., posting a photo), but can include other activity (e.g., a friend being tagged in a public photo, a friend commenting on a photo by another friend). Typically, many of these items score poorly or are removed according to rules to avoid repetitiveness (e.g., several items about friends commenting on the same photo), but some may score well.

4. Even in the same platform there are numerous channels for algorithmic ranking and recommendation. Public discourse often has in mind a single feed (e.g., Facebook News Feed, TikTok's series of videos), but these same items might be delivered to users via multiple feeds† and via email or mobile push notifications, with these likewise subject to optimization.[19] Established platforms are often ranking and recommending many types of items in many different formats. For example, many platforms suggest accounts to follow or friend based on numerous signals and with the aim of getting new users to make, e.g., engaging and varied connections.[20] Many platforms also have personalized search functionality in many places, including some that might seem mundane or invisible (e.g., autocompletion of friends' names when tagging them in photos).

Algorithmic transparency and impact: Evidence and challenges

There is substantial interest in characterizing "algorithmic amplification" — what content is given greater reach than it would have with some baseline algorithm — and the broader impact of algorithms. The title of this hearing refers to "dangerous algorithms" and posits associated harms. I find it quite plausible that, in particular cases, algorithmic ranking in social media is the proximal cause of specific harms; likewise,

*This is sometimes described as the "slate recommendation" problem, including in research from YouTube and others.[18] Facebook managers describe this as part of a "contextual pass" that, among other things, implements poster and content-type diversity rules.[11]

†For some time starting in 2011, Facebook had both its News Feed and a second, chronological feed (Ticker) that updated in real-time; both were visible simultaneously when using Facebook on a computer.

in particular cases, algorithmic ranking is the proximal cause of specific benefits and the absence specific harms. But how would we know whether aggregate effects of algorithmic ranking and recommendation in social media are positive or negative? To be clear, I do not think there is a social-scientific consensus here in favor of some simple baseline ranking (e.g., chronological) over the status quo.

Ideally, we would like to have rigorous, quantitative evidence about the effects of algorithmic ranking. While there presently is not a large body of evidence, one hope is that algorithmic transparency and other efforts would enable expanding this evidence. Here I review ways we can learn about algorithmic ranking and recommendation, in the process summarizing what limited evidence* we do have and highlighting some of the challenges in using simplistic comparisons of algorithms.[22]

I see three common ways we can learn about the effects of algorithms: (a) querying the algorithm with different inputs, (b) comparing outputs (i.e. rankings) of different algorithms on the same inputs (i.e. content inventory), and (c) conducting randomized trials assigning users (whether content producers or consumers) to different algorithms.

First, and most minimally, we may be able to **see how the algorithm ranks different items**; we can provide a variety of items (perhaps systematically varying some of their characteristics) and see the output. To some degree, this is commonly done by marketers (whether in commercial, public interest, or political campaigns) probing social media platforms in attempts to optimize their own reach by trying numerous variations on the content and timing of what they post. This is also an approach that has been used successfully to identify, e.g., disparate error rates in commercial computer vision systems.[23]† One challenge here is that one needs relevant samples of items — and perhaps the ability to generate systematic variations on them — to run through the algorithm. In the context of computer vision, there are available corpuses of images and creating a new sample of images is possible since they can be provided to the algorithm in standard formats. However, in the context of social media, external researchers may have little access to a distribution of items and many different signals are used in the ranking (i.e., the left column of Figure 1), many of which have only some proprietary format.

Second, we may be able to **compare how different algorithms rank the same inventory** for the same user. Academic researchers have, for example, set up Twitter accounts that emulated some archetypal real users and compared how news content is displayed in the "Home" (algorithmically ranked) versus "Latest Tweets" (approximately chronological) view; one study with eight artificial accounts found that the algorithmic ranking resulted in less exposure to external links as a

*Further afield, there are detailed studies of predictive systems making biased and harmful decisions in, e.g., health care,[21] a setting where both quite different regulatory considerations apply and some of the challenges described above as less severe.

†A variation on this approach would also use access to the algorithms' code itself. This might seem like a substantial advantage, but often the complete algorithm is complex enough — having potentially billions of parameters — that this may neither facilitate human understanding (even by the engineers building them) or readily enable external researchers or auditors to run a variation on the algorithm.

whole, including links to news.[24][*]

While these kinds of comparisons are an important tool, they typically do not tell us about "algorithmic amplification" writ large. These comparisons tell us about how items would be ranked if — for a moment and for one account only — the algorithm was changed; they do not typically tell us about what would happen if the algorithm were changed for a longer period of time and for many people or everyone.

Consider an example. Say you are connected on social media (e.g., are friends on Facebook, follow them on Twitter) to a relative with political views very different than your own and who posts a lot of political content. Due to algorithmic ranking and recommendation you might not see a lot of those posts, as the platform accurately predicts you will not engage with them nor would you say, if asked, that they helped you be informed or feel connected; rather you see a few of those posts, but you do see most of their posts about family and fishing. This might seem like a clear case of a "filter bubble" whereby the algorithm is causing you to be exposed to less cross-cutting content than you would under a simple chronological ranking.[†] However, the truth can be a bit more complex than that. Under chronological ranking you might initially see so many undesirable political posts from this person that you choose to unfollow or unfriend them. This could result in eventually seeing less cross-cutting content than you would under the algorithmic ranking.[‡]

More generally we can note that some of the effects of algorithmic ranking and recommendation of content occur by causing people to change their behavior including the formation, maintenance, and dissolution of their social network ties.[32] It can also affect the timing and duration of use of social media. Both of these then determine the longer-run consequences for what items they see. This process is illustrated in Figure 2.

Third, platforms can conduct **randomized trials comparing different ranking algorithms**. This is a key tool by which they optimize these algorithms with respect to their goals, and it is also used to gain further understanding of these complex systems.[33–35]

Randomized trials show that ranking choices can indeed matter. In the lead up to the 2012 US presidential election, a routine ranking experiment at Facebook randomly assigned over 1 million Americans to an algorithm that boosted the ranking of "hard news" from established news outlets; a preliminary analysis concluded that this increased political knowledge, altered policy preferences, and increased voter turnout compared with the status quo ranking.[36] Note, however, that the baseline here is not chronological ranking; in fact, the novel algorithm only moderately differed from the status quo, as it only affected a small fraction of items.[37][§] Thus, while this study provides evidence that, in a broad

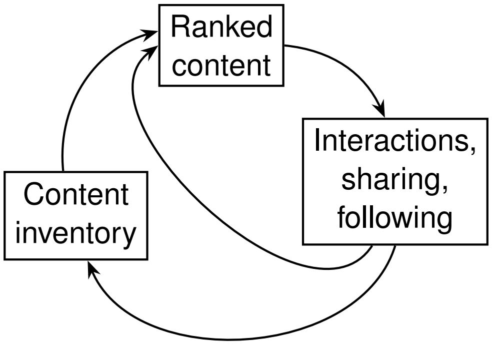
[*] These studies are often limited — by platform policies and enforcement — in the number of artificial accounts they can use to probe these algorithms.

[†] While studies by researchers at Facebook[25] document a reduction in cross-cutting political content from the set available from friends to that displayed in the feed, these researchers cautioned[26] that this was not immediately informative about what would happen in a counterfactual world with a chronological feed.

[‡] Perhaps in contrast to conventional wisdom, whether such reduced exposure to cross-cutting content is harmful for society is not always so clear,[26–28] further highlighting that the conclusions here are not so obvious. I am far from alone in arguing that some common claims associated with the "filter bubble" idea lack evidence.[29–31]

[§] With respect to some of the challenges described above and below, this is actually an advantage of this study, as we might expect less dramatic adjustments by users and publishers in response to such a change.

Figure 2: A simplified feedback loop in algorithmic ranking in social media. The content available to a user (i.e. their inventory*) depends on what accounts they follow. Even in the case of chronological ranking, what they see depends on their behavior (via, e.g., timing of use). Users change their behavior when the content they are shown changes (e.g., unfollowing other accounts, spending more or less time on platform).*



sense, *ranking matters*, it says less about how something approximately like the social media status quo compares with, e.g., a chronological ranking.

There are a small number of published studies directly comparing algorithmic ranking with simple heuristic ranking by chronology or overall popularity. These are largely in settings (e.g., music and podcast recommendations) removed from much of the public discussion about social media.* However, recently leaked documents from Facebook include a preliminary analysis of a test in which some users were randomly assigned to chronological ranking.[39,40] While we lack important details about this test, and a single such A/B test is not definitive, this reveals how, in the absence of ranking, metrics measuring exposure to likely "bad" content (e.g., spam, items people propogate and then delete) increased dramatically — as did these users' time spent on Facebook and the revenue attributed to them.† Users in chronological ranking also did not seem to like what they were seeing, as they hid many more items. As noted by the author of the internal report,[40] as with the comparisons described above, it is possible some of these effects are only short-run and they might reverse over the longer run as users remove friends, leave groups, or lower their expectations for Facebook and visit less.

A recent article reports on a similar, longer-running experiment at Twitter, in which some users were randomly assigned to chronological ranking.[41] These authors summarize their results as indicating that

*For example, Spotify conducted a test of their podcast recommendations in which some users received recommendations personalized based on their listening, while others received recommendations personalized based only on their demographics.[38] The more personalized group listened to somewhat more podcasts, but on average listened to a lower variety of categories of podcasts.

†This highlights that the status quo ranking was clearly not optimized to simply maximize short-run time spent or revenue.

politicians in general receive "algorithmic amplification" in that they get more reach among users in the status quo ranking compared with chronological ranking; across several countries, right-leaning parties are said to receive more amplification. While this long-running experiment avoids some of the challenges described above (e.g., these users have had time to adjust who they follow), it is still not obviously informative about what would happen in the counterfactual world where everyone has chronological ranking.[*]

To see an important remaining challenge here, consider how politicians or their social media managers would respond if all Twitter users have chronological ranking. Presently, they have tuned their Twitter activity to the status quo, but if suddenly everything was chronological, they would change their behavior.[†] Accounting for this kind of response from users — and especially from strategic, professional users of a platform — is key to assessing what is truly being amplified by an algorithm. Highly-transparent rankings and recommendations can be easier for various actors to "game".[‡] I would not be surprised if key results from these experiments might disappear or reverse when accounting for these responses.

Platforms clearly regard these feedback loops, spillovers across users, and adaptive, strategic behaviors as important, with several industry and industry–academic teams working on methods to better quantify these effects.[35,38,44] Thus, assessing what a ranking algorithm is amplifying is not a trivial task that platforms have simply neglected. And we may be substantially mislead by assessments of algorithmic amplification that simply compare two rankings of the same content or even compare consumption by users randomized to different algorithms.

Qualitative evidence

In the absence of evidence from quantitative studies, we might look to more qualitative evidence. Commentators frequently point to cases where algorithmic recommendation of, e.g., content or groups is an antecedent of some harm. But it is unclear how much of these outcomes to attribute to specific features of social media, and algorithmic ranking in particular. One useful comparison is that clearly inflammatory content and misinformation can very much go viral in the absence of such ranking, with similarly terrible immediately consequences. For example, viral rumors spread on the group messaging service WhatsApp[§] — which lacks algorithmic ranking — have been the proximal cause of lynching and other violence in India.[46] This hardly proves that algorithmic ranking does not matter, but it certainly highlights the difficulty of

[*] The Twitter researchers refer to these challenges briefly, citing some of my work[42] and noting that the experiment does not "provide unbiased estimates of causal quantities of interest".[41]

[†] Sophisticated social media managers would be optimizing this quantitatively, but even individual users (including politicians) would receive feedback via the amount of likes and retweets their different posts get or may view basic aggregate reach statistics provided by Twitter.

[‡] Rankings by recent overall popularity are also often subject to strategic behavior as well. For example, we observed groups coordinate their Twitter activity to make dozens of political hashtags appear on Twitter's trending topics list during the 2019 Indian general election.[43]

[§] In our study of images shared in politics-related WhatsApp groups in India, images identified as misinformation by journalists made up around 13% of all image shares.[45]

credibly attributing harms to algorithmic ranking based on these kinds of observations alone.

This absence of general, credible evidence about effects of algorithmic ranking and recommendation partially reflects the limited ability of external researchers, journalists, and others to probe these systems. Thus, we may expect that efforts to promote algorithmic transparency and to enable producing reports on algorithmic impact using internal, proprietary data could facilitate the creation of better evidence. But this absence also reflects some fundamental challenges, some of which I have elaborated on here, to learning about effects of interventions in these complex environments — even with access to all of the proprietary data.

Policy implications

It is beyond the scope of my testimony and my expertise to make comprehensive policy recommendations. Instead I just highlight a few implications of the preceding discussion.

- Policy-making should not presume that requiring or promoting the use of largely unfiltered chronological feeds in social media will be beneficial. There could be benefits from resulting transparency (e.g., requiring platforms to offer choice may make it easier for external researchers to learn about the algorithms), and perhaps to individual consumers. But there may be other substantial costs, including substantial increases in spam and misinformation without corresponding direct improvements for consumers.

- Minimal versions of algorithmic transparency — such as requiring disclosure of ranking algorithms — may not be sufficiently informative to yield many benefits. Depending on the values of policymakers, this might suggest these efforts are not worthwhile, or that a broader and more substantial version of transparency, involving disclosure of substantially more data, is needed.

- Algorithmic ranking and recommendation are ubiquitous, often appearing in many distinct channels on the same platform. Policymakers may want to consider whether a particular regulation should apply to all of these channels, especially if compliance requires substantial effort and user-facing controls for each channel (e.g., a prominent icon) and if some channels are already close substitutes for each other.

- Algorithmic ranking and recommendation are not limited to the decision of whether to display and item and in what order, as items can be aggregated or displayed more or less prominently. Expanding this repertoire of user interfaces can be an important site of innovation. Transparency mandates that neglect this may (a) fail to be informative about important decisions by platforms and/or (b) impede innovation in new ways of presenting information and media.

- There are existing (e.g., TikTok, YouTube) and potential social media platforms where there is no obvious "default" ranking. Policies that would enshrine comparisons to a particular baseline when defining "algorithmic amplification" may lead to absurd conclusions for many platforms.

- Naive definitions of "algorithmic amplification" can be misleading because they do not account for the reactions of consumers and (often quite strategic) producers, which can be dramatic. Alongside other problems, this may complicate efforts to use such definitions to establish moral or legal responsibility.[47]

- External researchers (whether academics, journalists,[48] or government officials) depend on the ability to probe these systems. Various data-collection methods often opposed by platforms (including scraping public content, participants consenting to automated contribution of their data,[49,50], confirming participants have stopped using a service,[51,52] and sophisticated audit studies[24,53–55]) remain central to their work. If policy-makers want rigorous external research — whether on foreign interference in US elections[56] or on effects of algorithms — to be possible, they can protect the right to employ these methods, rather than, e.g., the Computer Fraud and Abuse Act discouraging such work.[57]

- Some of the best evidence I have drawn on comes from A/B tests (i.e. randomized trials) conducted by platforms. These are regarded as a "gold standard" for learning about cause-and-effect relationships, and they are a key tool for understanding effects of changes to algorithmic ranking, as well as many other key decisions. Policies and rhetoric discouraging their use could leave scientists and the public less informed — as well as make it harder for the platforms themselves to make good decisions.[58]

- Arrangements whereby platforms share data with external researchers while protecting people's privacy can be beneficial for science and the public. However, there remains substantial uncertainty about what methods for privacy protection satisfy some international privacy regulations,[59] perhaps discouraging more of such sharing. In some cases, privacy regulations have apparently discouraged platforms from even retaining detailed data that would be useful for assessing algorithmic impact.[56]

About me

I am a researcher and educator working at the intersection of communication technologies and topics of broad societal interest, including political participation, public health, and commerce. In addition to doing empirical research, I work on methodology for learning about cause and effect relationships (i.e. causal inference), particularly when people are affected by others' behaviors, as in much of social life, but especially in social media.

I am faculty at the Massachusetts Institute of Technology (MIT), where I am currently the Mitsubishi Career Development Professor, an associate professor in the Marketing group of the Sloan School of Management, and affiliated faculty at the Institute for Data, Systems & Society in the Schwarzman College of Computing, including its Center for Statistics and Data Science. I teach analytics in our professional degree programs and research methodology to PhD students in multiple programs. My academic research has been published in peer-reviewed journals and proceedings spanning several fields. I studied at Stanford University, resulting in five degree including my PhD.

While currently an academic, I have broad knowledge of the Internet industry and state-of-the-art practices there. I was previously a scientist and consultant at Facebook, where I worked on feed, messaging, advertising, survey methods, and tools for randomized experiments. Before that I worked at Nokia and Yahoo, where I likewise worked in research, studying and designing social media. I co-organize an annual conference that involves substantial participation by data scientists and managers from the Internet industry, and I frequently have industry experts as guest speakers in my classes at MIT.

Acknowledgements

I am grateful to comments from several people, including Jennifer Allen, Daniel Björkegren, Tom Cunningham, Andrey Fradkin, Marta Franco, David Rand, and Johan Ugander. Thanks to Alex Kantrowitz for allowing me to review a key document[40] used in his reporting.[39] Responsibility for this document is mine alone.

Disclosures

Here I note some relevant potentially competing interests. My full set of potentially competing interests are listed on my website. I am a consultant to Twitter. My research has been funded by Amazon, Boston Globe Media, Facebook, IBM, the US Air Force (via a subcontract from Lincoln Laboratory), and The World Bank. A conference I co-organize has been sponsored by Amazon, Facebook, and Netflix. I conduct some of my research via both data use agreements with unnamed firms (e.g., retailers, Internet companies, and fitness companies) and by advising research by collaborators, often students, who are employees or consultants of other firms (e.g., Facebook). I have significant financial interests in Amazon, GoFundMe, Google, and Salesforce.

References

1. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.
2. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. doi:10.1038/nature11421
3. Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., & Fowler, J. H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US Presidential Election. *PLOS ONE*, *12*(4), e0173851.
4. Aral, S. & Nicolaides, C. (2017). Exercise contagion in a global social network. *Nature Communications*, *8*(1), 1–8.
5. Holtz, D., Zhao, M., Benzell, S. G., Cao, C. Y., Rahimian, M. A., Yang, J., . . . Sowrirajan, T. et al. (2020). Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences*, *117*(33), 19837–19843.
6. Aral, S. (2021). *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health—and How We Must Adapt*. Currency.
7. Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013). Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 21–30).
8. Mosseri, A. (2021). Shedding more light on how Instagram works. *Instagram*. Retrieved from https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works
9. Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.
10. Backstrom, L. (2016). Serving a billion personalized news feeds. In *Proceedings of the ninth ACM International Conference on Web Search and Data Mining* (pp. 469–469). Retrieved from https://www.youtube.com/watch?v=Xpx5RYNTQvg
11. Lada, A., Wang, M., & Yan, T. (2021). How does news feed predict what you want to see? *Facebook Newsroom*. Retrieved from https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/
12. Smith, B. (2021). How TikTok reads your mind. Retrieved from https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html
13. Oremus, W. (2016). Who controls your Facebook feed. *Slate*. Retrieved from http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html
14. Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2019). Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, *14*(2), 495–519.
15. Obeng, A. & Bakshy, E. (2020). Preference learning for real-world multi-objective decision making. In *ICML 2020 Workshop on Real World Experiment Design and Active Learning*.
16. Eckles, D., Kizilcec, R. F., & Bakshy, E. (2016). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, *113*(27), 7316–7322.
17. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30.
18. Wilhelm, M., Ramanathan, A., Bonomo, A., Jain, S., Chi, E. H., & Gillenwater, J. (2018). Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 2165–2173).
19. Zhao, B., Narita, K., Orten, B., & Egan, J. (2018). Notification volume control and optimization system at Pinterest. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1012–1020).

20. Su, J., Kamath, K., Sharma, A., Ugander, J., & Goel, S. (2020). An experimental study of structural diversity in social networks. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 661–670).
21. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.
22. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... Jackson, M. O. et al. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.
23. Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR.
24. Bandy, J. & Diakopoulos, N. (2021). Curating quality? How Twitter's timeline algorithm treats different types of news. *Social Media + Society*, *7*(3), 20563051211041648.
25. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.
26. Bakshy, E. & Messing, S. (2015). *Exposure to ideologically diverse news and opinion, future research*. Retrieved from https://solomonmg.github.io/post/exposure-to-ideologically-diverse-response/
27. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.
28. Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–70.
29. Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). *Avoiding the echo chamber about echo chambers*. Knight Foundation.
30. Hosanagar, K. & Miller, A. P. (2020). Who do we blame for the filter bubble? On the roles of math, data, and people in algorithmic social systems. In K. Werbach (Ed.), *After the Digital Tornado: Networks, Algorithms, Humanity* (pp. 103–121). Cambridge University Press.
31. Bail, C. (2021). *Breaking the Social Media Prism*. Princeton University Press.
32. Berman, R. & Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, *39*(2), 296–316.
33. Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World Wide Web* (pp. 283–292).
34. Peysakhovich, A. & Eckles, D. (2018). Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference* (pp. 699–707).
35. Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., ... Coey, D. et al. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, *21*(1), 20–35.
36. Messing, S. (2013). *Friends that matter: How social transmission of elite discourse shapes political knowledge, attitudes, and behavior*. Doctoral dissertation, Chapter 7. Stanford University.
37. Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539.
38. Holtz, D., Lobel, R., Liskovich, I., & Aral, S. (2020). Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489*.
39. Kantrowitz, A. (2021). Facebook removed the News Feed algorithm in an experiment. Then it gave up. *Big Technology*. Retrieved from https://bigtechnology.substack.com/p/facebook-removed-the-news-feed-algorithm
40. Anonymous. (2018). *What happens if we delete ranked News Feed?* Facebook. Part of "The Facebook Files".
41. Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2021). Algorithmic amplification of politics on Twitter. *arXiv preprint arXiv:2110.11010*.
42. Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1).

43. Jakesch, M., Garimella, K., Eckles, D., & Naaman, M. (2021). Trend alert: A cross-platform organization manipulated Twitter trends in the Indian general election. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2). doi:10.1145/3479523

44. Karrer, B., Shi, L., Bhole, M., Goldman, M., Palmer, T., Gelman, C., . . . Sun, F. (2021). Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3106–3116).

45. Garimella, K. & Eckles, D. (2020). Images and misinformation in political groups: Evidence from Whatsapp in India. *Harvard Kennedy School Misinformation Review*.

46. Dwoskin, E. & Gowen, A. (2018). On WhatsApp, fake news is fast — and can be fatal. *Washington Post*. Retrieved from https://www.washingtonpost.com/business/economy/on-whatsapp-fake-news-is-fast--and-can-be-fatal/2018/07/23/a2dd7112-8ebf-11e8-bcd5-9d911c784c38_story.html

47. Keller, D. (2021). Amplification and its discontents: Why regulating the reach of online content is hard. *Journal of Free Speech Law*, *1*, 227–268.

48. Matt, F., Stern, J., Barry, R., West, J., & Wells, G. (2021). How TikTok's algorithm figures out your deepest desires. *Wall Street Journal*. Retrieved from https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

49. NYU researchers were studying disinformation on Facebook. the company cut them off. (2021). *NPR*. Retrieved from https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f

50. Brinberg, M., Ram, N., Yang, X., Cho, M.-J., Sundar, S. S., Robinson, T. N., & Reeves, B. (2021). The idiosyncrasies of everyday digital lives: Using the human screenome project to study user behavior on smartphones. *Computers in Human Behavior*, *114*, 106570.

51. Brynjolfsson, E., Collis, A., & Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences*, *116*(15), 7250–7255.

52. Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–76.

53. Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22.

54. Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., Sandvig, C. et al. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human–Computer Interaction*, *14*(4), 272–344.

55. Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, *12*(1), 1–10.

56. Aral, S. & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858–861.

57. Sandvig v. Sessions. (2018). District Court, District of Columbia.

58. Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colorado Technology Law Journal*, *13*, 273.

59. Altman, M., Cohen, A., Nissim, K., & Wood, A. (2021). What a hybrid legal-technical analysis teaches us about privacy regulation: The case of singling out. *Boston University Journal of Science and Technology Law*, *27*, 1.