BROOKINGS

Report

@elonmusk and @twitter: The problem with social media is misaligned recommendation systems, not free speech

Justin Bullock and Anton Korinek Wednesday, May 18, 2022

Abstract

Elon Musk's negotiations to take over Twitter have put a spotlight on the meaning of free speech and content moderation in the age of social media, which are powered by AI-based algorithms called "recommendation systems." They have also served as a reminder of the stark tension between the public role of media companies in connecting and informing our society and their private role in financing their activities and earning a profit. Recommendation systems that maximize user engagement are really effective in the latter private role, but potentially at the expense of undermining the public goods that media companies have traditionally produced for society. We argue that it is crucial to align the objectives programmed into recommendation systems with broader societal values. This would enable social media companies to play a more beneficial role in society.

Introduction

On April 14, 2022, Elon Musk offered to buy Twitter in an all-cash deal worth $43 billion after having accumulated a significant stake in the company in the weeks prior. The immediate response (on social media) seemed to break along traditional political lines, with conservatives hailing the deal as a win for free speech and liberals expressing concerns about Musk's approach to content moderation. For his part, Musk polled followers (more than once) on Twitter to get their thoughts on improving the platform. He also stated a set of initial priorities for changing how Twitter is governed. A recent Reuters article puts Musk's focus this way: "Musk, who calls himself a free speech absolutist, has

criticized Twitter's moderation. He wants Twitter's algorithm for prioritizing tweets to be public and objects to giving too much power on the service to corporations that advertise." Musk and Twitter have engaged in negotiations about the bid for weeks now, with Twitter's board formally accepting the offer on April 25. However, more recently, Musk cast doubt on the deal by citing that Twitter may have more fake accounts than previously acknowledged, in what may be an attempt to renegotiate or walk away from the deal.

In this brief, we analyze the role of algorithms and human values in moderating the content of social media platforms by investigating the tension between the public role that media companies play in our society and the private economic incentives that they face.

The public good of curating information

Since the beginning of humanity, the technologies that we have employed to exchange information have played a central role in our societies. The invention of language allowed early humans to exchange information and develop human culture, setting us apart from all other species. The invention of writing allowed for novel ways of organizing large groups of humans and accelerated human development. The efficient collection, storage, processing, and distribution of information is crucial for the functioning of our modern society.

In today's world, there is so much information that we need mechanisms to curate it—that is organize, select, and present it—allowing for the dissemination of broader stories and ideas, narratives and news. A curated flow of information allows our society to remain connected, to adapt to new challenges, to learn, and to function more efficiently.

Traditional media organizations have long played an essential role in organizing and curating information, dating back to the first newspaper in the early 17th century, followed by the radio and national and cable television in the 20th century. Each of these media—brought to life by emerging technology of the time —worked differently and affected how information is transmitted, but they all had in common that they played the role of curating information for a significant part of the population, synchronizing the flows of information to large groups of people. For example, the use of the earliest printing presses helped spur on the Protestant Reformation. The radio was used to spread information and

galvanizing propaganda during WWII by the Allied and Axis powers, helping to shape morale and the national narratives. The television brought with it televised presidential debates, changing the sets of factors and tools campaigns had to consider. For example, the televised debates between Kennedy and Nixon influenced the narrative around the election and the public perception of the candidates. The pivotal public role of media organizations assigns them both great power and great responsibility.

Having a well-functioning media landscape that effectively disseminates information is an important public good that generates large positive externalities for society. Given these externalities, some societies fund media organizations from the public purse, e.g., the BBC in the U.K., in order to fulfill the media's core function of information provision to the public. However, public ownership of media also carries the risk of being leveraged by autocrats.

In the U.S., most media organizations are private, for-profit organizations. This structure generates its own set of risks—the main good that media companies have traditionally provided is the public good of curating information, but the value of that good is difficult to capture because it is what economists call non-rival and effectively non-excludable. In other words, news can easily be shared with others. The revenue that media companies can capture is thus only a fraction of the public goods value that they contribute to society. Media companies are largely financed by advertising—an ancillary activity that cross-subsidizes the collection, curation, and distribution of news. By the same token, the market value of media companies reflects only that small sliver of the value created through advertising and subscriptions.

The effects of AI on the media landscape

Technology has impacted the media landscape in two fundamental ways in the past two decades. First, advertising migrated from traditional media to online advertising, undermining the main revenue source of traditional media and leading to job cuts in news rooms and closures, especially among local newspapers. This was not because traditional news media no longer created value for society – it was because technological change

undermined their (ancillary) revenue model. For example, this is what enabled Jeff Bezos to acquire the Washington Post—one of the leading newspapers in the US—for a fire-sale price of $250 million in 2013.

Second and more recently, the internet and the AI revolution enabled the rise of a new type of media company, the social media company. Like traditional media, social media still plays the role of curating information and therefore has correspondingly important effects on information flows, with a similar potential to impose large externalities on society. However, social media differ from traditional media in very important ways when it comes to content creation and selection. Traditional media is mostly a one-way communication system where editors, journalists, and advertisers decide on the content that is created and delivered. The consumer does not typically play an active role in creating and selecting content. The situation is entirely different with social media. Social media relies on user-generated content and curates information algorithmically. Instead of journalists, social media companies employ algorithms called recommendation systems to moderate content and decide which information to serve users. The over-arching goal programmed into these recommendation systems is to maximize "user engagement"—that is, to maximize how much time the user spends on the platform and interacts with the content, which correlates closely with ad revenue generated.

Modern recommendation systems use factors such as the personal characteristics of a user, the relationship of a user with the authors of posts, the type of content, and the popularity of posts and their recency together with minute details about the preferences of their users. The user profiles created by these systems are based on past engagement with the platform to predict what posts are most likely to capture a user's attention. The more a user engages with the platform, the better the recommendation system can personalize the content that it serves to better "engage" the user, inducing them to spend additional time, see more advertisements, and in turn generate more revenue for the social media company. This creates a feedback loop that allows the algorithm to maximize user engagement ever more effectively.

The externalities imposed by recommendation systems

What is missing from the feedback loop is accounting for the public goods nature of the information provision of media companies. In traditional media, journalists and editors brought a long-standing set of professional norms and ethical standards to their work, used their discretion and interests to select and curate the topics that they considered of interest to society, and fact-checked and carefully selected their sources. These norms act like a form of self-regulation that seeks to ensure the reputation and quality of the content that is presented. They are reinforced in professional schools, trainings, and media organizations. The decisions that need to be made about what is newsworthy are complex and contextual to the nuances of the setting. In other words, they are judgment calls made by human professionals about which stories to deliver to the reader and viewer.

The situation is very different for social media. To begin with, much of the content on social media is generated by individual users, and advertisers, as well as by traditional media companies. This content is not verified and often is not even intended to be informative. But decisions still have to be made about which users receive which content. In other words, the content still requires some mechanism for curation and moderation for distribution of the information. For social media, the role of editors is supplanted by recommendation systems, which have no inherent understanding of truth or the complexity of human society. As a result, social media sites have struggled with a range of undesirable outcomes that have imposed massive negative externalities on society, as documented, for example, in the so-called Facebook Files, from creating echo chambers and increasing political polarization to serving misinformation and toxic content to minors.

Recognizing these shortcomings and feeling the public pressure to rein in such externalities, social media companies have attempted to complement the metric of maximal user engagement with other mechanisms of content moderation. In recent years, they have hired tens of thousands of content moderators to tag, remove, or block content that is violent, sexually explicit, or offensive in a way that crosses red lines established by the social media companies. Although we applaud such efforts, the results have been mixed. Moreover, they have also led to significant drawbacks. For one, removing posts

opens social media companies to the critique of censorship. Second, and more importantly, a focus on policing whether a given piece of content crosses certain red lines does not address the root of the problem: as long as posts do not cross the red lines laid out by human content moderators, they are fair game for the recommendation system to propagate. And in fact, it is frequently precisely the posts that come closest to those red lines that incite strong reactions and generate maximal engagement. Such recommendation systems thus act like a human moderator in a debate who would pick up only on the most outrageous comments and push other participants to focus attention and react to those—hardly a good recipe for a healthy conversation. What is really needed is that recommendation systems pursue a different metric than short-term user engagement—a metric that includes what type of content fulfills media companies' responsibility to society, given the powerful effects they have on society's information flows.

Elon Musk and the governance of Twitter

From this perspective, Elon Musk's takeover of Twitter gives us reason for both hope and concern. Musk seems to understand the powerful role of social media in our society's information flow. He observed that buying Twitter "is not a way to make money," although his bankers and investors will expect a return on their investment. But the crucial question is this: is he willing to take into account and internalize the large externalities that social media organizations like Twitter impose on society?

Musk has repeatedly emphasized the importance he places on free speech. Who can argue with the value of free speech? It is hard, actually, to overstate the value of a society where speech is free and protected from abuse by the government and other organizations. Freedom of speech is closely related to our collective desire for freedom of thought, something that must also be protected at the societal level. But what does Musk mean by claiming he is a "free speech absolutist?"

In a narrow sense, "free speech" on social media platforms is simply the ability to post content. However, this is an almost meaningless concept: if the platform's algorithm does not serve a post to other users, it is unlikely to be seen – it is like speaking into the void.

We generally support letting any human post content on Twitter (subject to certain relatively narrow red lines) but not necessarily the amplification of that content.

There is another naïve and dangerous sense in which free speech can be interpreted—to let anybody post any content and have automated recommendation systems, with no understanding of the responsibility of social media to society, decide whether to serve that content to other users. We hope that Musk recognizes that the objectives that social media companies program into their recommendation systems are separate from whether speech is free or not. Like all media organizations—both traditional and social media—Twitter has a responsibility to promote information flows in a way that benefits society rather than further polarizing us.

There is no such thing as a "neutral" way in which recommendation systems can operate. Every decision they make to serve one post over another is an active decision. Whether a recommendation system is programmed to focus on short-term "user engagement" by promoting outrage or whether it is programmed to improve social harmony is not a question of free speech—it is a question of what objectives the system has been programmed to pursue. Recommendation systems that maximize user engagement have clear biases towards certain types of content that may appeal to our impulses but are undesirable for society and thereby drown out other types of content that are more desirable. K. Eric Drexler puts it the following way:

"Social media today is degraded by the influence of [misinformed] and unsourced content, a problem caused (in part) by the cost of finding good information and citing sources, and (in part) by fact-indifferent actors with other agendas. Well-informed replies are relatively costly and scarce, but mob-noise and bot-spew are abundant."

One of Musk's stated goals is to make public the algorithm behind Twitter's recommendation system that decides which tweets to prioritize. This would allow the public to inspect the algorithm more closely and could open up a public debate on what private and social metrics we want the recommendation system to pursue. However, it could also make it easier for unscrupulous actors to manipulate the algorithm.

Another avenue to improve the effects of modern social media on society's information flows would be to give users more explicit choices about the content that they are served. For example, Twitter could add buttons such as "Inform me" to let users request high-quality content, "Challenge me" to let users see content from different viewpoints within our society, or "Entertain me" to explicitly let users request lower-quality information.

Finally, another stated goal of Musk is to take further steps to authenticate users that are real persons. A significant amount of content on Twitter is generated or amplified by bots, i.e., automated tools that copy content or endorse content in order to manipulate information flows, sometimes at the behest of foreign powers. Authenticating users and prioritizing content produced by their accounts could serve this objective.

Social media recommendation systems and the broader challenge of AI alignment

Ultimately, the debate on what goals the recommendation systems of social media platforms should pursue mirrors a broader debate in the field of AI governance on how to make sure that AI systems pursue the goals that we want them to pursue – the so-called AI alignment problem. There are two aspects of this alignment problem. The direct alignment problem is concerned with the question of whether an AI system does what its operator wants it to do. The social alignment problem or AI governance problem is concerned with whether the AI system's conduct is beneficial for society at large. In the case of Twitter, its AI-based recommendation system seems to be in direct alignment – it is very effective at keeping users engaged on the platform. Twitter has also attempted to address issues of social alignment, but with mixed results so far – it still imposes negative externalities on society, externalities akin to "polluting" the information space of the public dialogue.

Traditional media – and together with them, the professional norms of journalists and editors – evolved over many decades. By contrast, social media have sprung up over a much shorter time span. We are only starting to recognize the fundamental ways in which they reshape the information flows within our society and present new governance challenges. Our proposals on how to respond to these challenges are therefore modest. We believe that the following steps would be beneficial for social media companies to take in order to aid progress towards reducing negative externalities and generating greater positive externalities:

1. As Musk also suggests, social media companies should be required to be transparent about the objectives that their recommendation systems optimize for, allowing for public inspection and debate on how the systems affect our society.
2. These objectives should not only focus on user engagement but also account for broader social objectives. This, of course, comes with its own significant challenges in implementation. While it is not clear yet what the best governance mechanisms are for encouraging social media companies to limit the harms done by their massively scaled impact, it is nonetheless important that we find pathways to require these large companies to internalize the externalities they impose.[1] Governance tools for consideration here include regulation of the goals and capabilities of recommendation algorithms, such as limiting their use to influence minors, and the use of oversight boards to help set and enforce norms against predatory advertising with practices such as pooling data across platforms.
3. At least for the foreseeable future, humans should be in the loop when it comes to deciding what content is recommended, providing the recommendation systems with feedback to improve the sensitivity to human norms for communication and information exchange.
4. It would be helpful for users to be better authenticated to improve the quality of content through increased trust that the content was created and posted by a fellow human. This could be accomplished, for example, by attempting to eliminate bots more rigorously and authenticating that users are indeed humans.

More broadly, the contentious debate that followed Musk's takeover bid for Twitter has highlighted both the importance of the structure of our social media landscape for society and the role of AI-based recommendation systems within that structure. Public policy needs to pay greater attention to the role of such recommendation systems. We do not advocate that governments should unilaterally decide how content is moderated. However, as one of us has argued in a related report, regulators need new powers to oversee the deployment of advanced AI systems, including the recommendation systems that underlie social media companies. Society is still in the process of determining what social objectives to assign to recommendation systems.

Musk is aware of the dangers of unaligned AI systems, calling them humanity's "biggest existential threat". In 2015, he donated $10 million to the Future of Life Institute to support research on how to "keep AI beneficial to humanity." As a founder and former board member of OpenAI, the leading AI research lab working on advanced language models, he is also aware of the power of modern AI systems to analyze, aggregate, evaluate, and even generate information in human language—a capability that complements Twitter's trove of content in potentially disquieting ways. We hope that he will continue his commitment to "keep AI beneficial to humanity" by recognizing and living up to the social responsibility that owning one of the world's leading social media organizations entails—even if this comes at the expense of reducing the virality of some of his own tweets.

The Brookings Institution is financed through the support of a diverse array of foundations, corporations, governments, individuals, as well as an endowment. A list of donors can be found in our annual reports published online here. The findings, interpretations, and conclusions in this report are solely those of its author(s) and are not influenced by any donation.

Footnotes

1. 1 Keller and Leerssen (2020) provide an up-to-date view of content moderation practices by social media companies.