Testimony for "Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content"

Senate Committee on Homeland Security and Governmental Affairs

Cathy O'Neil, CEO of ORCAA

October 28, 2021

Introduction

When we think about a user's experience on social media platforms, the raw ingredients are the pieces of content that people post, but it's the algorithms which decide who sees what and when (the "recommendation algorithm"). There are also separate algorithms (the "filter algorithms") which try to determine whether a given piece of content contains hate speech or is otherwise not allowed by policy.

The main goals of my testimony will be trying to explain, in very concrete terms, first what an algorithm is, second what a recommendation algorithm is and how recommendation algorithms lead people to become more extreme, and finally what a filter algorithm is and why we should be skeptical of their efficacy. I will conclude with an explanation of algorithmic audits and how they might be useful for understanding the harmful impact of social media platforms.

What is an algorithm?

How do you decide what to wear in the morning? You look in your closet at the available clothes and, depending on your memories and how you think your day will unfold, you decide what to wear. If you need to look professional, you'll choose a different outfit than if you're merely trying to be maximally comfortable.

You are using a "getting dressed" algorithm. Let me explain.

An algorithm needs two ingredients to do pattern matching: historical data and a definition of "success". The data can be simply memories stored in your head or digitized records stored on a computer. The definition of success can likewise be personal feelings of "being comfortable" or "being professional" or they can be mathematical formulas that computers can understand.

The algorithm sifts through the historical data provided to it, looks at examples of "success" as defined, and identifies patterns from the past that distinguished the successful cases from the

others. The algorithm learns from historical data what was successful in the past and predicts similar things to be successful in the future. In the examples of getting dressed, you could have a bad memory of wearing a particular pair of pants that ended up being uncomfortable, which would lead you to discard this choice (if the goal today is to be comfortable). If a computer is doing this pattern matching, the fancy math behind these techniques is able to find subtle, complicated patterns that people often can't. In fact, they sometimes find patterns that aren't even explainable or understandable to people.

Two important points: first, the definition of success really matters. If you want to look professional day after day, you'll end up wearing very different outfits than if you are optimizing for comfort. Second, whomever decides what "success" looks like for a powerful algorithm is actually wielding an enormous amount of power. Algorithms deployed by businesses are given a definition of success that primarily optimizes to profit. That might not be - and quite often, isn't - what is best for the rest of us; imagine an algorithm that tells us to wear uncomfortable pants every single day.

What is a recommendation engine?

Recommendation engines are a specific kind of algorithm focused on the task of making a personalized suggestion that will appeal to you -- say, a Facebook group to join, or a person to follow on twitter or Instagram. In the language of the first section, the Facebook newsfeed takes as historical data everything you've ever done on Facebook, and defines success as "keeping you on Facebook for as long as possible." Other social media platforms do very similar things, and for the same reason: the longer you are on their platform, the more you click on ads, which is how they make money. So they optimize for profit.

Behind the scenes, recommendation engines rely on hundreds of categories that represent real-life topics or interests. For instance, categories might include "baseball," or "crafts," or "cute animals," or "the stock market," or "politics." Each person gets a score for every category, and higher scores indicate a higher level of interest in that topic. So a person like me, who's a knitting fanatic but doesn't care about sports besides baseball have a high score for "crafts" and for "baseball" but a low score for "basketball" and "football."

Similarly, each piece of content gets a score in every category, representing its relevance to that topic. So a box score from last night's Red Sox game will have a high score for "baseball" and a low score for "crafts."

To a recommendation engine, each person and each piece of content is represented by a long list of scores. The algorithm then makes matches between people and content based on these lists. Basically, it serves up content whose scores closely mirror the user's scores.

These scores are constantly updated as people see and interact with (or ignore) content. If a user is shown the Red Sox box score and comments on it or shares it, their score for “baseball” will increase and their other scores will proportionately diminish. Similarly, if a knitting tutorial video gets shared and liked by many users with high “crafts” scores, then the video’s “crafts” score will increase. In this way the user is “teaching” the algorithm what their scores are by every single action or inaction.

How do recommendation engines create rabbit holes?

Because people with high “baseball” scores are more likely to be served content with high “baseball” scores, they have more opportunities to push their “baseball” score *even higher* by interacting with that content. Similarly, since I love knitting already, any time a cashmere yarn store advertises to me on social media and I click on the irresistible ad, the algorithm will take note and set my “knitting” score higher, which means I’ll be more likely to be shown knitting related content, especially ads, in the future.

This feedback loop gets supercharged when you introduce “viral” content that is outrageous or sensational. Whether it is happy or sad, inspiring or infuriating, the more sensational it is, the more likely it is to elicit a reaction from a user. Whether that reaction is a share/repost or a comment condemning the content, it counts as engagement and updates the scores, causing it to be shown even more. The more “viral” the content, the faster this happens. Facebook changed its algorithm to further boost viral content, and the result was more divisiveness and more extreme content.

Filter algorithms

Certain kinds of viral content -- the kinds that promote violence, or contain hate speech -- are exactly what is causing harm to individuals and society. Facebook understands this completely and has been spending the last few years trying to mitigate the harms by creating filter algorithms. Think of these as somewhat more sophisticated but similar to keyword searches you might run on your email account in order to find a specific email from a few weeks or months ago: you’d search by the name of the person who wrote to you and a word or two that was special in the conversation snippet that you remember.

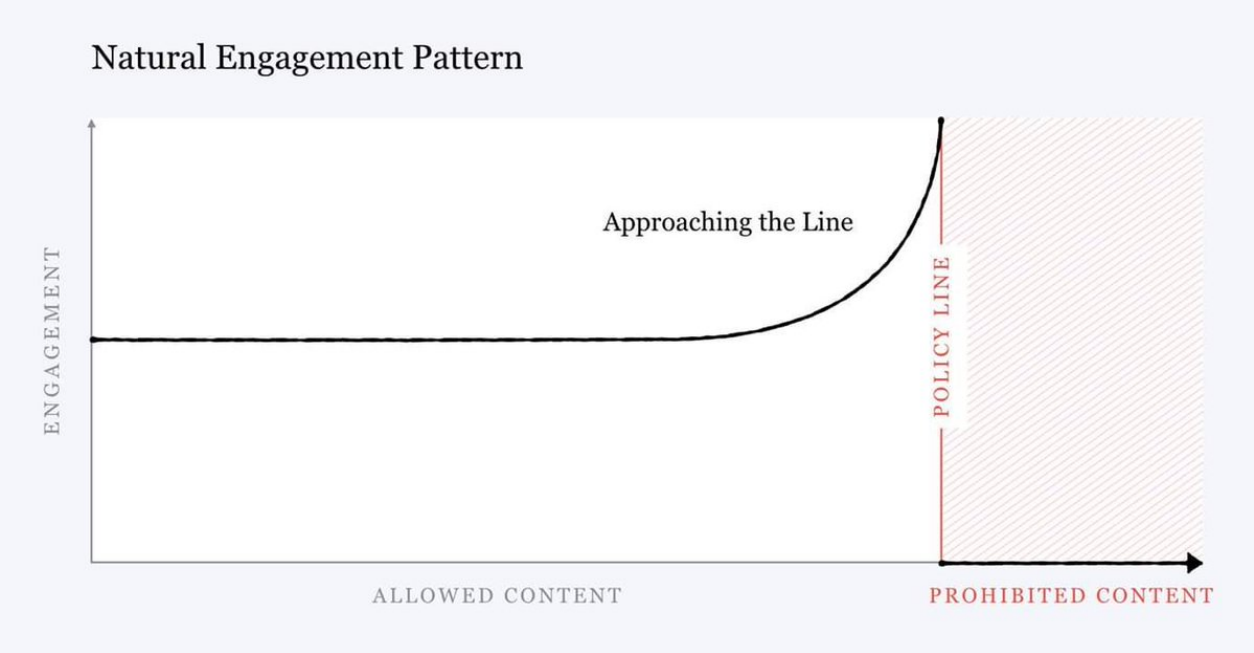
Similarly, filter algorithms look for keywords that are associated with hate speech, misinformation, or conspiracy theories. They are trained on historical pieces of content that have been labeled definitively prohibited in the past. So if someone posted the same thing again, it would get caught. If they posted something that’s almost the same, it might get caught. But if

they posted something that had the same message but said in a novel way, it wouldn't get caught.

The problem for social media platforms is that there are people who get paid to bypass their filter algorithms with hateful content, or misinformation, or conspiracy theories. So it's an army of living humans, who are clever and strong willed, against an algorithm that cannot keep up.

Facebook's internal research shows its AI moderation tools successfully catch a mere 2 - 5% of all prohibited content. If you think of the filter algorithm as a net that is supposed to catch a certain type of fish, you should imagine that almost all of the fish manage to swim through the net without being threatened by the net at all.

Even in the context of 2-5% of content that is actively being scrutinised, the news is troubling. This graph from a 2018 letter by Mark Zuckerberg described the problem: content to the right of the "policy line" is bad and must be prohibited; but content just slightly to the left of the policy line is exactly what the platform wants, since it garners the most engagement.



Later in the same letter, Zuckerberg explains that Facebook will begin to identify "borderline" content and "downweight" it (i.e., show it less) so that this curve bends down instead of up. But even if they did that, it's still only going to apply to the small fraction of harmful content that they observe.

Conclusion: we need to audit these algorithms

At my company ORCAA, an algorithmic audit starts with the simple question: For whom could this fail? This centers the audit around stakeholders: the people whose lives could be impacted by the algorithm, for better or worse. Our audit process involves talking directly to stakeholders to understand their concerns.

In order to audit a social media platform's recommendation algorithm, then, we would need a complete list of stakeholders. But with billions of users, this is impossible. Nobody can anticipate all the specific subgroups of users that could have distinct concerns. I doubt even the most diligent auditor would have predicted that the Rohingya Muslims in Myanmar would need to be considered as a distinct stakeholder group.

If we can't audit the algorithm as a whole, what can we do? We can start by fixing some specific groups of stakeholders and focusing on particular harms. Then we can demand social media companies provide evidence -- or perhaps to give over raw data that allows regulators or other researchers to produce evidence -- that their products do not cause these harms to these stakeholders.

It's not enough for them to ignore the harm, or to research it in such a minimal way that they can deny it's really a problem when that research is leaked.

Questions you might want to ask me

- What is an algorithm?
- What types of algorithms are used by social media platforms?
- How algorithms push users down rabbit holes?
- What can be done?