Nakyazze Pricilla Research Discussion 4

June 23, 2025

How to counter the radicalizing effects of recommender systems or ways to prevent algorithmic discrimination.

To counter the radicalizing effects of recommender systems and prevent algorithmic discrimination, we must acknowledge the core problem, recommendation algorithms prioritize engagement over well-being, optimizing for clicks, watch time, or ad revenue rather than informed choice or social harmony.

As Renee DiResta and Zeynep Tufekci argue, platforms like YouTube and Facebook often push users toward extreme content because their algorithms are trained to maximize watch time, not truth or diversity. This leads to polarization (YouTube's push toward more extreme videos), misinformation amplification (Pinterest disinformation boards, Facebook Groups).

BuzzFeed recently reported that Facebook Groups nudge people toward conspiratorial content, creating a built-in audience for spammers and propagandists. Follow one ISIS sympathizer on Twitter, and several others will appear under the 'Who to follow'.

The systems don't actually understand the content; they just return what they predict will keep us clicking. That's because their primary function is to help achieve one or two specific key performance indicators (KPIs) chosen by the company.

In the 1950s, Solomon Asch performed a well-known set of experiments. When answering in the presence of a group of confederates who agreed on incorrect answers, 25% of participants conformed to the incorrect consensus values.

Sanjay Krishnan, Jay Patel, Michael J. Franklin, and Ken Goldberg from the Department of Electrical Engineering and Computer Sciences, UC Berkeley, proposed a methodology to learn, analyze, and mitigate the effects of social influence bias in recommender systems.

UC Berkeley's research shows how exposure to others' opinions (e.g., average ratings) distorts personal judgments, reinforcing majority views and homogenizing preferences a danger for diverse or counter-majority perspectives.

Can we make the internet's recommendation engines more ethical? And if so, how?

Diversity aware recommendation, where algorithms should not only optimize for similarity but also for diversity of viewpoints or content.

Projects such as Project Redirect (Google Jigsaw), where when a user searches for extremist content, they are redirected to counter-narratives or de-radicalizing content. Rather than offer up more violent content, the approach of that recommendation system is to do the opposite.

Let users adjust their feed settings, filter low-quality content, and view content from outside their echo chamber. Giving people more control over what their algorithmic feed serves up is one potential solution. Twitter, for example, created a filter that enables users to avoid content from low-quality accounts. Not everyone uses it, but the option exists.

Recommendations should be explainable. For example, users should know why something is being recommended. The way Netflix does it is they say, 'Because you watched' vs. opaque auto-play queues. They also say, "We think you will love this based on past selections."

Platforms should use 'nudges' to promote healthy behaviors. The same way schools should put apples in front of chips for students, trustworthy news should be prioritized and clickbait demoted. This is considered a soft paternalism approach that preserves freedom of choice but guides toward better decisions.

Social influence bias, as shown by Krishnan et al., distorts feedback loops in recsys.

Use triplet rating analysis to detect bias, then apply correction models in two ways: train the correction on all triplets, including ones that did not change, to get a correction that we can then apply to all ratings in the post-learning phase. The second way is to estimate the probability that a rating has changed, and if that probability is above a threshold (For example 50 percent), we can apply the correction. With the second way, the correction model is only trained on those triplets where the initial rating is different from the final one.

Estimate users' true ratings without social pressure and use these corrected ratings to train fairer recommendation models.

For large inventories, cluster items into domains (e.g., politics, lifestyle, religion), and train bias correction models per domain.

Algorithms must be subject to external audits, just like financial systems.

Platforms should disclose how their recommender systems work, what KPIs they optimize for, and their known side effects.

Conclusion

Moderation is often framed as censorship, but unmoderated recommendation is also a form of bias. It's selective amplification.

The debate isn't 'moderation vs. freedom', but which values the algorithm promotes by default. If we don't actively shape these systems, clickbait, extremism, and misinformation win by design.

To prevent recommender systems from becoming "The Great Polarizer," we must rethink their purpose, designing for informed diversity, not addictive similarity. Letting users steer their own algorithmic experience. Holding platforms accountable for the real-world impacts of their black-box designs.

Recommendation engines are not neutral. They are engines of influence. We must build them as if that power matters because it does.

The need to rethink the ethics of recommendation engines is only growing more urgent as curatorial systems and AI crop up in increasingly more sensitive places. Local and national governments are using similar algorithms to determine who makes bail, who receives subsidies, and which neighborhoods need policing. As algorithms amass more power and responsibility in our everyday lives, we need to create the frameworks to rigorously hold them accountable—that means prioritizing ethics over profit.