Regulating Recommendation Algorithms On Social Media: Challenges And Strategies

LAWS428: Law and Emerging Technologies

Optional Research and Writing Assignment

2,997 words

I Introduction

Across the world, billions of hours are spent on social media each day. Whoever decides what we get shown there holds enormous power. Today, it is a given that complex algorithms decide what to show each user. But those algorithms are relatively recent; only in 2015 did Facebook's algorithm start to use information on how long people hover on a particular item.[1]

The algorithms which have become so central to the business model of social media giants have since come under scrutiny as it became evident they can cause severe harm.

First, they undermine the free trade of ideas which is so fundamental to democracy through 'algorithmic audiencing.' In other words, they algorithmically interfere with who reads what online. It is "akin to large-scale social engineering."[2] An example of how this might be a problem is the finding that across six countries Twitter's feed-ranking algorithm amplified posts from elected officials on the political right more than posts from those on the left.[3]

More dramatically, they increases polarisation and hate. Facebook's own research found that the recommendation algorithms encourage extremism.[4] And as former employee and whistleblower Frances Haugen describes it:[5]

[1] Greg Kumparak "Facebook Now Cares About How Long You Look At Stuff In Your News Feed" (12 June 2015) TechCrunch <https://techcrunch.com/2015/06/12/facebook-now-cares-about-how-long-you-look-at-stuff-in-your-news-feed/>.

[2] Kai Riemer and Sandra Peter "Wrong, Elon Musk: the big problem with free speech on platforms isn't censorship. It's the algorithms" (18 May 2022) TheConversation < https://theconversation.com/wrong-elon-musk-the-big-problem-with-free-speech-on-platforms-isnt-censorship-its-the-algorithms-182433>.

[3] Rumman Chowdhury and Luca "Examining algorithmic amplification of political content on Twitter" (21 October 2021) TwitterBlog <https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent>, cited in Tom Simonite "Europe's New Law Will Force Secretive TikTok to Open Up" (4 May 2022) Wired <https://www.wired.com/story/tiktok-transparency-dsa-europe/>.

[4] Taylor Hatmaker "Facebook hits pause on algorithmic recommendations for political and social issue groups" (31 October 2020) TechCrunch <https://techcrunch.com/2020/10/30/facebook-group-recommendations-election/>.

[5] Frances Haugen "Statement at the United States Senate Committee on Commerce, Science, and Transportation Sub-Committee on Consumer Protection, Product Safety, and Data Security" (4 October 2021) <https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49>, cited in Sachin Holdheim "Regulating Content Recommendation Algorithms in Social Media" (paper presented at the Digital Platform Regulation Conference, Yale University, May 2022).

> The result has been a system that amplifies division, extremism, and polarization — and undermining societies around the world. In some cases, this dangerous online talk has led to actual violence that harms and even kills people. In other cases, their profit optimizing machine is generating self-harm and self-hate — especially for vulnerable groups, like teenage girls. These problems have been confirmed repeatedly by Facebook's own internal research.

Clearly, there is a serious need for regulation to address these harms. This report discusses the challenges to regulating recommendation algorithms and strategies for overcoming those. There are many uses for artificial intelligence algorithms. This report focusses on recommendation algorithms only, and solely on their use in deciding which content appears first in social media feeds. They will be referred to as Content Recommendation Algorithms (**CRAs**).

II Challenges

Three broad challenges will be discussed. First and most pertinent, the difficulty in knowing what is going on in a CRA, in order to know how to regulate. Second, the challenge of weighing up the costs and benefits of CRAs. There is significant uncertainty around the costs of regulating. Third, the challenge of regulation without putting regulatory power into the wrong hands.

A Understanding what you are regulating

CRAs are known for their opacity. Because regulators do not understand how the algorithm works, it is difficult to know what risks they need to protect against and how regulation should be structured and enforced. There are many reasons for this incomplete understanding.

For one thing, CRAs will often be trade secrets,[6] and social media companies tend to hold information about them close to their chest.[7] Beyond the obvious reasons why they would want to do this, there is the very legitimate security concern that the more people who have access to CRAs, the more people who can learn to exploit them.[8] Those people may

[6] Frank Pasquale *The Black Box Society* (Harvard University Press: Boston, 2016), cited in Karen Yeung "Algorithmic regulation: A critical interrogation" (2017) 12 Regulation & Governance at 26.

[7] Renée DiResta et al "It's Time to Open the Black Box of Social Media" (28 April 2022) ScientificAmerican < https://www.scientificamerican.com/article/its-time-to-open-the-black-box-of-social-media/>.

[8] Will Knight "Elon Musk's Plan to Open Source the Twitter Algorithm Won't Solve Anything" (27 April 2022) Wired <https://www.wired.com/story/twitter-open-algorithm-problem/>.

wish merely to gain prominence, but equally they may wish to promote material with malicious intent. For instance, destabilise another country by spreading conspiracy theories.

But the complexity of machine learning algorithms means that even with access to complete information, regulation is not simple. For one thing, there is not one single algorithm at play: "decisions are the result of many different algorithms that perform a complex dance atop mountains of data and a multitude of human actions."[9] And for another, algorithms work differently depending on the data fed to them (by the many users on Twitter, for instance):[10]

> [Twitter's machine learning models] cannot be inspected like regular code; they need to be tested in an environment that replicates the real world as closely as possible. The models also change rapidly in the real system, in response to a constant flow of new data, user behavior, and input from moderators. This would quickly make them an unreliable source of information.

So problems with the use of algorithms are not obvious merely from looking at the coding of an algorithm. Part of the opacity of machine learning algorithms is inherent in the 'rewards system' used. It means that a company is able to specify the goals of the system, but will not necessarily know how the AI will achieve those goals, including the 'instrumental goals' it might deem necessary to achieve them.[11] Companies themselves may not know how their algorithms are making decisions. The ability to understand their own algorithms further diminishes as programming staff come and go, and as the algorithm self-learns.[12]

So there is significant opacity to CRAs. This raises a particular challenge when trying to regulate how an algorithm treats a certain type of content. That's because the algorithm does not aim to promote a *type* of content. It merely attempts to achieve goals, such as increased engagement, often by identifying what a user engages with and showing more and more specific or extreme examples of that. The machines themselves generally do not have the capacity to understand the content, only to predict what will keep users clicking. For instance, the algorithm would not know the difference between "suggesting

[9] Above, n 8.

[10] Above, n 8.

[11] Fiona Seal "Regulating Artificial Intelligence: A Critical Analysis of Technology Law's Gordian Knot in the New Zealand Context" (LLM, University of Otago, 2021).

[12] Henry Flood "Adequate Protection: An Examination Of Transborder Data Protection Standards In The European Union" (LLM Thesis, University of Otago, 2017).

duckie balloons and serving up extremist propaganda."[13] This makes regulation that would prohibit amplifying only a certain type of content much less feasible. At least, compliance with such a regulation would require human input to categorise content, which would be highly resource-intensive.

The opacity of CRAs also raises challenges regarding *how* one regulates. Ordinarily, a regulator would set a standard for what a tool can or cannot do, and if it breached that standard then either an individual affected or an auditor would take action against the wrongdoer. However, individuals are rarely going to know when they have been discriminated against or their decisions influenced by a CRA.[14] It is difficult for a body to audit CRAs without black-box access (access to a simulated environment which mirrors the algorithm in its real world environment), and even black-box access will not allow regulators to predict how the algorithm will operate in future environments. Then, when bringing action against a wrongdoer, it is difficult to show causation in the sense courts are used to because "they are based on patterns and correlations between data points, rather than on a causal or explanatory theory of behaviour, and are continuously reconfigured in light of past input and output data."[15] Even a company may not know how the algorithm came to a particular decision.

B Uncertainty surrounding costs of regulation

Regulators are well versed in cost-benefit analyses. Costs considered include not only explicit costs of regulation but also the lost utility of the tool that is being restricted. The unique challenge when it comes to CRAs is the uncertainty around the costs of regulation: what would social media look like today if CRAs were prohibited or restricted?

Very little research exists on the risks of prohibiting or limiting the use of CRAs, and that may be because we have not experienced an online world without CRAs. Those social media companies which do not use CRAs (Telegram is a good example) do not have the same level of engagement as social media giants like Facebook, Twitter, TikTok and YouTube and so are not good comparisons.

[13] Renee Diresta "Up Next: A Better Recommendation System" (11 April 2018) Wired <https://www.wired.com/story/creating-ethical-recommendation-engines/>.

[14] Above, n 6.

[15] Above, n 6.

Some argue that eliminating CRAs will undermine the very things that regulation is trying to protect: democratic free speech, reducing hate speech, an empowered public. 'Dataism' is the belief that humans can no longer usefully sift through the large amount of information 'out there' and it underpins much of the justification for CRAs. CRAs might encourage a better free trade of ideas because they enable worthy ideas from people with small followings to be amplified; people do not have to be well-known in order to be heard.[16] In terms of empowering the public, algorithms may have been instrumental to some degree in fueling movements such as the Arab Spring and Black Lives Matter.[17] Although clearly biased, Facebook's Vice President of Global Affairs claims that without the algorithm, users would see more hate speech, misinformation and harmful content.[18]

So eliminating CRAs altogether may be costly, but there may also be risks to attempting to retain just the 'good bits' of CRAs. For instance, recommending more and more specific cat videos to cat-lovers, but not recommending extremist content to those who show an interest in that. Whether it is appropriate to allow algorithmic amplification of only some types of content is hotly debated. One side of the debate argues that this kind of regulation amounts to social engineering, or at least choice architecture, and it interferes with freedom of speech, potentially running afoul of the First Amendment in the US.[19] The other side of the debate draws a distinction between free speech and 'free reach.' It argues that while you have a right to express thoughts, you do not have a right to an algorithmic amplification those thoughts through recommendations to larger audiences.[20]

C Keeping power from the wrong hands

Then there is the challenge of who should be entrusted with regulating. Access to data about social media activity and control over the algorithms used confers huge power on

[16] Will Oremus et al "How Facebook shapes your feed" (26 October 2021) TheWashingtonPost <https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/>.

[17] Will Oremus "Lawmakers' latest idea to fix Facebook: Regulate the algorithm" (12 October 2021) TheWashingtonPost <https://www.washingtonpost.com/technology/2021/10/12/congress-regulate-facebook-algorithm/>.

[18] Julia Cherner "Facebook exec says company will make itself 'more transparent'" (11 October 2021) ABCNews <https://abcnews.go.com/Politics/facebook-exec-company-make-transparent/story?id=80498312>.

[19] Above, n 17.

[20] Renee Diresta "Free Speech Is Not the Same As Free Reach" (30 August 2018) Wired <https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/>; above, n 17.

social media giants. That warrants oversight or regulation, but it is equally concerning giving governments a share in that power. In March this year, TikTok created a Russian echo chamber so that Russian users could not see content posted by non-Russian channels.[21] Edward Snowden's disclosures in 2013 revealed that the United States' NSA conducted secret mass surveillance by collecting information from Facebook and YouTube servers, among others.[22] More recently, a hacking group speculated to be Chinese state-sponsored conducted espionage on Uyghurs through Facebook.[23] Evidently, governments do not have a good track record for using the power of social media responsibly. Procedural legitimacy is equally as important as substantive legitimacy in regulation.[24] To that end, the power to regulate should rest with an entity which has good authority. Although governments have good political and legal authority, their track record makes one question whether they are the right body to take on such power.

III Strategies

I have discussed three key challenges: opacity, the uncertainty as to the costs of regulation and the risk of putting regulatory power into the wrong hands. I now move on to discuss possible strategies for addressing those.

D Transparency

To support any kind of regulation, a key priority is addressing the opacity problem. So far regulators have proposed to do this by: requiring vetted researchers access to the an

[21] Salvatore Romano et al "Tracking Exposed Special Report: TikTok content restriction in Russia" (15 March 2022) TrackingExposed < https://tracking.exposed/pdf/tiktok-russia-15march2022.pdf>; cited in Tim Simonite "TikTok's Black Box Obscures Its Role in Russia's War" (28 March 2022) Wired < https://www.wired.com/story/tiktok-algorithm-russia-war/>.

[22] Barton Gellman Laura Poitras "U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program " (7 June 2013) TheWashingtonPost <https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html>.

[23] Mike Dvilyanski "Taking Action Against Hackers in China" (24 March 2021) AboutFB <https://about.fb.com/news/2021/03/taking-action-against-hackers-in-china/>; Ellen Nakashima "Facebook disrupts China-based hackers it says spied on Uyghur Muslim dissidents and journalists living outside China, including in the U.S" (24 March 2021) TheWashingtonPost <https://www.washingtonpost.com/national-security/china-espionage-uyghurs-facebook/2021/03/24/7f2978d2-8c38-11eb-a6bd-0eb91c03305a_story.html>.

[24] Roger Brownsword and Morag Goodwin *Law and the Technologies of the Twenty-First Century : Text and Materials* (Cambridge University Press, 2012) at 48.

Application Programming Interface and subjecting platforms to external audits.[25] Critiques of this approach include that even with black-box access, auditing is not a straightforward process. One must still figure out how to design an audit to enforce it, whether the audit imposes a high performance cost on the platform, and how the audit affects the content that the platform is incentivized to filter.[26] Research is being done to address these obstacles, but regulators would likely need to work alongside technical experts in order to make sense of it. Then, even if auditing identifies a bias, it will not always be clear why the bias is occurring or how to eliminate it (as described earlier).

Further proposals include: requiring companies to submit regular reports to governments about changes to algorithms or developments in their operation and requiring increased government oversight during the development stage.[27] Of course this encounters the challenge of handing power to governments discussed earlier, and pushback from companies themselves.

E Effectively prohibit CRAs

In terms of regulation itself, one option is to have a strong regulatory tilt towards precaution. If regulators subscribed to the strong version of the precautionary principle they might put the onus on social media giants to prove that CRAs will not cause harm to society.[28] Due to their complex and opaque nature, such 'proof' would likely be near-impossible, and so effectively kill CRAs. People would tend to be in favour of this approach if they believe that: any level of algorithmic amplification interferes with the free marketplace of ideas; that regulation of only some types of content is too close to social engineering or is not achievable; or that it is not safe to give any one body the power to decide what should not be amplified. Critics of this approach argue that CRAs strengthen the free marketplace of ideas.

F Prudential pluralism

One option to dealing with the uncertainty about what the risks and benefits of CRAs are is to facilitate individual choice over whether to use them. This could happen by requiring platforms to offer a feed not manipulated by algorithms, something the European Digital

[25] Above, n 3.

[26] Sarah Cen and Devavrat Shah "Regulating algorithmic filtering on social media" (Paper presented at the 35th Conference on Neural Information Processing Systems, 2021).

[27] Sachin Holdheim, above n 5.

[28] For a description and defence of the strong precautionary principle: Noah Sachs "Rescuing the Strong Precautionary Principle from its Critics" (2011) U.Ill.L.Rev. 1285.

Services Act and the US Filter Bubble Transparency Bill look to do.[29] Facebook, Instagram and Twitter offer this currently,[30] but given how powerful default settings are, they likely need to be made more accessible or deliberate.[31] Regulators could also look to require an explanation be provided to users on how the CRA decides on their recommendations.[32] For instance, the EU's General Data Protection Regulation introduces a right to "meaningful information about the logic involved" in algorithmic decisions, its significance and consequences.[33]

A proposal that goes even further suggests that platforms be required to offer a 'sandbox' to users whereby they "would be able to see their recommended content ordering as if they were users with different characteristics, but the same underlying content set."[34] While users would not be able to alter their feed, it would dramatically increase understanding of how CRAs shape perspective and the kinds of presumptions CRAs make about a person. This runs into the aforementioned security risk of 'bad actors' learning to manipulate algorithms.

G Restrict CRAs

If regulators do not think it necessary to effectively prohibit CRAs nor appropriate to leave it up to individuals, there are some options for merely setting standards with which CRAs must abide.

Currently, it appears platforms such as Facebook would not be liable if a user posted something which attracted civil or criminal liability (e.g., hate speech or defamatory comments). In the United States for instance, website platforms generally have immunity

[29] Above n 3; above, n 17.

[30] Isobel Asher Hamilton "How to switch your Facebook feed to a chronological timeline" (10 January 2022) TheBusinessInsider <https://www.businessinsider.com/facebook-social-media-switch-feed-chronological-timeline-2021-11?op=1>; John Kennedy "The button that will put your Instagram feed in chronological order" (29 March 2022) PopSci+ < https://www.popsci.com/diy/how-to-make-instagram-feed-chronological/>; J. D. Biersdorfer "Putting Your Twitter Feed Back in Chronological Order" (21 March 2016) TheNewYorkTimes <https://www.nytimes.com/2016/03/22/technology/personaltech/putting-your-twitter-feed-back-in-chronological-order.html>.

[31] Nick Babich "The Power of Defaults" (13 April 2017) UXPlanet < https://uxplanet.org/the-power-of-defaults-992d50b73968>.

[32] Sachin Holdheim, above n 5; in the context of targeted advertising: Tokeley and Stace, eds. *Consumer Law in New Zealand* (forthcoming, 2022).

[33] Regulation (EU) 2016/679 (General Data Protection Regulation), art 13(2)(f), cited in Flood, above n 12.

[34] Sachin Holdheim, above n 5.

with regard to third-party content.[35] One option is to make platforms liable when they *amplify* through recommendations (as opposed to just host) posts which violate certain civil rights.[36] Regulators could prohibit amplification of medical misinformation or particular political material. One instance of this approach is YouTube's Project Redirect. That feature works by steering people who search for certain keywords away from violent extremist propaganda and "toward video content that confronts extremist messages and debunks its mythology."[37] The challenge with this type of regulation is difficulty of compliance. It may result in platforms erring on the side of caution, which disproportionately affects marginalised communities,[38] or effectively killing CRAs altogether. Further, there is great controversy around only allowing the amplification of certain content. As discussed earlier, some argue this amounts to choice architecture and undermines the free marketplace of ideas.

An option for a regulatory standard which would alter algorithms at a more fundamental level is to change the performance measures of the algorithm. Rather than optimise for engagement, it could take into account 'truthfulness,' or specifically, "a quality indicator derived from a combination of signals about the content, the way it's disseminated (are bots involved?), and the authenticity of the channel, group, or voice behind it."[39] Alternatively, regulators can require algorithms not discriminate based on certain characteristics.

Setting standards by which CRAs must abide is likely to be most popular among regulators. To some extent it preserves commercial freedom of platforms, while aiming to prevent social harm.

H Structural solutions

Several creative solutions have been suggested which go to the structure of regulation or of the market.

One approach is to aim regulation at protecting user data.[40] This would reduce the capability of CRAs while retaining freedom as to how programmers develop the

[35] Communications Decency Act 1996 (US), s 230.

[36] Above, n 17.

[37] The Youtube Team "Bringing new Redirect Method features to YouTube" (20 June 2017) YoutubeOfficialBlog <https://blog.youtube/news-and-events/bringing-new-redirect-method-features/>.

[38] Above, n 17.

[39] Sachin Holdheim, above n 5; above, n 13.

[40] Above, n 17.

algorithm. Alternatively, aim regulation at the efforts platforms make, rather than the outcomes. Mary Ann Franks, an American legal scholar, suggests liability for any company that "manifests deliberate indifference to unlawful material or conduct."[41] That would go towards mitigating social harm without creating a sudden and impracticable compliance burden for companies. However, that might enable companies to do what looks good to outsiders, regardless of whether it is effective. Twitter did this when it looked to remove likes to appease academics and governments, despite internal research reportedly showing it to be minimally effective.[42]

One very elegant solution is to have third parties create algorithms and enable users to choose between those.[43] That would remove the profit incentive of platforms and would prevent governments taking over too much control. Further, it would introduce a strong incentive to make algorithms transparent to users, in order to compete with the other algorithms on offer.

China has been quicker and bolder in their regulation than other countries. This might not come down to structural differences in the approach regulation, but rather cultural differences.[44] The Western world is tied up by the notion that companies have a right to freedom of innovation, and that free markets with minimal interference produce the best social outcomes. Privacy is seen as an individual choice rather than a collective good. Meanwhile, China has approached this as a simple welfare and stability problem and taken a more interventionist and collective approach. That approach has resulted in some very robust regulation.

IV Conclusion

The algorithms which recommend what billions of people see for hours of each week are extremely influential. Regulating them to prevent harm is challenging because it is not clear what harm they might cause, nor what benefit they currently confer, and because the regulator themself would acquire dangerous power. Strategies to overcome this uncertainty include taking a strong precautionary stance and effectively prohibiting CRAs, leaving it up to individual choice, restricting what CRAs can do, or using more

[41] Gilad Edelman "Congress Takes Aim at the Algorithms" (2 December 2021) Wired <https://www.wired.com/story/congress-takes-aim-at-algorithms-section-230-reform/>.

[42] Above, n 17.

[43] Above, n 17.

[44] Above, n 41.

creative structural solutions. Regardless of the approach, regulators should be aware of the Western bias towards individualism.