18

Filter Bubbles and Disinformation

Imagine that you're a technology executive who is unhappy with the stranglehold that a handful of companies have on how people receive information via ad-supported social media timelines, recommendations, and search engines. Your main issue with these 'big tech' companies is the filter bubbles, disinformation, and hate speech festering on their platforms that threaten a functioning non-violent society. Many of these phenomena result from machine learning systems that help the platforms maximize engagement and revenue. Economists call these considerations that extend beyond revenue maximization for the company and are detrimental to society *negative externalities*. According to your values, recommendation and search to maximize engagement are problems that should not even be worked on in their currently prevailing paradigm because they have consequences on several of the items listed in Chapter 14 (e.g. disinformation, addiction, surveillance state, hate and crime).

> "The best minds of my generation are thinking about how to make people click ads. That sucks."
>
> —Jeff Hammerbacher, computer scientist at Facebook

In recent months, you have seen an upstart search engine enter the fray that is not ad-driven and is focused on 'you,' with 'you' referring to the user and the user's information needs. This upstart gives you a glimmer of hope that something new and different can possibly break through the existing monopolies. However, your vision for something new is not centered on the singular user 'you', but on plural society. Therefore, you start planning a (fictional) search engine and information recommendation site of your own with a paradigm that aims to keep the negative externalities of the current ad/engagement paradigm at bay. Recalling a phrase that the conductor of your symphonic band used to say before concerts: "I nod to you and up we come," you name your site Upwe.com.

Does Upwe.com have legs? Can a search engine company really focus on serving a broader and selfless purpose? Many would argue that it is irrational to neither focus on solely serving the user (to make it attractive for paying subscribers) nor maximizing the platform's engagement (to maximize the

company's ad revenue). However, as you learned in Chapter 15, corporations are already moving toward broadening their purpose from maximizing shareholder value to maximizing the value for a larger set of stakeholders. And by focusing on the collective 'we,' you are appealing to a different kind of ethics: relationality instead of rationality. *Relational ethics* asks people to include considerations beyond themselves (which is the scope of rational ethics), especially their relationships with other people and the environment in determining the right action. One effect of relational thinking is bringing negative externalities to the forefront and mitigating an extractive or colonial mindset, including in the context of machine learning.[1]

So coming back to the original question: is Upwe.com tenable? Does your vision for it have any hope? In this chapter, you'll work toward an answer by:

- sketching the reasons why society is so reliant on the digital platforms of 'big tech,'
- examining the paradigm that leads to echo chambers, disinformation, and hate speech in greater detail, and
- evaluating possible means for countering the negative externalities.

18.1 Epistemic Dependence and Institutional Trust

As you're well aware, the amount of knowledge being created in our world is outpacing our ability to understand it. And it is only growing more complex. The exponential increase in digital information has been a boon for machine learning, but perhaps not so much for individual people and society. There is so much information in the modern world that it is impossible for any one person, on their own, to have the expertise to really understand or judge the truth of even a sliver of it. These days, even expert scientists do not understand the intricacies of all parts of their large-scale experimental apparatus.[2] Known as *epistemic dependence*, people have to rely on others to interpret knowledge for them. You've already learned about epistemic uncertainty (lack of knowledge) and epistemic advantage (knowledge of harms possessed by people with lived experience of marginalization) in Chapter 3 and Chapter 16, respectively. Epistemic dependence is along the same lines: obtaining knowledge you lack from people who possess it, trusting them without being able to verify the truth of that knowledge yourself.

The people from whom you can obtain knowledge now includes anyone anywhere at lightning speed from their messages, articles, blog posts, comments, photos, podcasts, and videos on the internet. Epistemic dependence no longer has any bounds, but the space of knowledge is so vast that it requires search engines and recommendation algorithms to deal with retrieving the information. And what is going on behind the scenes is almost never clear to the user of a search engine. It is something abstract and mysterious in the ether. Even if the seeker of knowledge were aware of an information retrieval algorithm's existence, which is typically based on machine learning, its workings would not be comprehensible. So not only do you have to trust the source and content of the knowledge, but also the

[1]Sabelo Mhlambi. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." Harvard University Carr Center Discussion Paper Series 2020-009, Jul. 2020.
[2]Matthew Hutson. "What Do You Know? The Unbearable Vicariousness of Knowledge." In: *MIT Technology Review* 123.6 (Nov./Dec. 2020), pp. 74–79.

closed-box system bringing it to you.[3] Nonetheless, people cannot entirely abdicate their epistemic responsibility to try to verify either the knowledge itself, its source, or the system bringing it forward.

From the very beginning of the book, the trustworthiness of machine learning systems has been equated to the trustworthiness of individual other people, such as coworkers, advisors, or decision makers. This framing has followed you throughout the journey of becoming familiar with trustworthy machine learning: going from competence and reliability to interaction and selflessness. However, when discussing the trustworthiness of the machine learning backing information filtering in digital platforms, this correspondence breaks down. To the general public, the machine learning is beyond the limits of their knowledge and interaction to such a degree that the machine learning model is not an individual person any longer, but an institution like a bank, post office, or judicial system. It is just there. Members of the public are not so much users of machine learning as they are subject to machine learning.[4] And institutional trust is different from interpersonal trust.

Public trust in institutions is not directed towards a specific aspect, component or interaction with the institution, but is an overarching feeling about something pervasive. The general public does not go in and test specific measures of the trustworthiness of an institution like they may with a person, i.e. assessing a person's ability, fairness, communication, beneficence, etc. (or even care to know the results of such an assessment). Members of the public rely on the system itself having the mechanisms in place to ensure that it is worthy of trust. The people's trust is built upon mechanisms such as governance and control described in Chapter 14, so these mechanisms need to be understandable and not require epistemic dependence. To understand governance, people need to understand and agree with the values that the system is working to align itself toward. Thus as you envision Upwe.com, you must give your utmost attention to getting the paradigm right and making the values understandable to anyone. Putting these two things in place will enable the public to make good on their epistemic responsibility. Remember from Chapter 15 that intervening on the paradigm is the most effective leverage point of a system and is why the focus of this chapter is on the paradigm rather than on tackling negative externalities more directly, such as methods for detecting hate speech.

18.2 Maximizing Engagement, or Not

So how can you get the paradigm and values right? There are many things that you can do, but the main one is to deprioritize engagement as the primary goal. Engagement or attention is often measured by a user's time on the platform and by their number of clicks. Maximizing engagement can lead to the extreme of the user becoming addicted to the platform.

---

[3]Boaz Miller and Isaac Record. "Justified Belief in a Digital Age: On the Epistemic Implications of Secret Internet Technologies." In: *Episteme* 10.2 (Jun. 2013), pp. 117–134.

[4]Bran Knowles and John T. Richards. "The Sanction of Authority: Promoting Public Trust in AI." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Mar. 2021, pp. 262–271.

> "When you're in the business of maximizing engagement, you're not interested in truth. You're not interested in harm, divisiveness, conspiracy. In fact, those are your friends."
>
> —Hany Farid, computer scientist at University of California, Berkeley

First, let's see how single-mindedly valuing engagement leads to the harms of echo chambers, disinformation, and hate speech. The end of the section will briefly mention some alternatives to engagement maximization.

18.2.1 Filter Bubbles and Echo Chambers

When a recommendation system shows a user only content related to their interests, connections, and worldview, they are in a *filter bubble*. But how do filter bubbles relate to maximizing engagement with a digital platform? This kind of curation and personalization keeps serving the user content that they enjoy, which keeps them coming back for more of the same. Pleasant and fun things attract our attention.

> "When you see perspectives that are different from yours, it requires thinking and creates aggravations. As a for-profit company that's selling attention to advertisers, Facebook doesn't want that, so there's a risk of algorithmic reinforcement of homogeneity, and filter bubbles."
>
> —Jennifer Stromer-Galley, information scientist at Syracuse University

In an *echo chamber*, a person is repeatedly presented with the same information without any differences of opinion. This situation leads to their believing in that information to an extreme degree, even when it is false. Filter bubbles often lead to echo chambers. Although filter bubbles may be considered a helpful act of curation, by being in one, the user is not exposed to a diversity of ideas. They suffer from *epistemic inequality*.[5] Recall from Chapter 16 that diversity leads to information elaboration—slowing down to think about contentious issues. Thus, by being in a filter bubble, people are apt to take shortcuts, which can lead to a variety of harms.

18.2.2 Misinformation and Disinformation

What are those fun things that attract us? Anything that is surprising attracts our attention.[6] There are only so many ways that you can make the truth surprising before it becomes old hat.[7] Permutations and combinations of falsehoods can continue to be surprising for much longer and thus keep a user more

---

[5]Shoshana Zuboff. "Caveat Usor: Surveillance Capitalism as Epistemic Inequality." In: *After the Digital Tornado*. Ed. by Kevin Werbach. Cambridge, England, UK: Cambridge University Press, 2020.
[6]Laurent Itti and Pierre Baldi. "Bayesian Surprise Attracts Human Attention." In: *Vision Research* 49.10 (Jun. 2009), pp. 1295–1306.
[7]Lav R. Varshney. "Limit Theorems for Creativity with Intentionality." In: *Proceedings of the International Conference on Computational Creativity*. Sep. 2020, pp. 390–393.

engaged on a platform. Moreover, people spread false news significantly faster on social media platforms than true news.[8]

> "Having constructed a technological apparatus that disseminates information instantaneously and globally without regard to its veracity, we shouldn't be surprised that this apparatus has left us drowning in lies."
>
> —Mark Pesce, futurist

Clickbait is one example of false, surprising, and attractive content that drives engagement. It is a kind of *misinformation* (a falsehood that may or may not have been deliberately created to mislead) and also a kind of *disinformation* (a falsehood that was purposefully created to mislead). In fact, 'big tech' companies have been found to finance so-called clickbait farms to drive up their platforms' engagement.[9]

> "Misinformation tends to be more compelling than journalistic content, as it's easy to make something interesting and fun if you have no commitment to the truth."
>
> —Patricia Rossini, communications researcher at University of Liverpool

Another type of disinformation enabled by machine learning is *deepfakes*. These are images or videos created with the help of generative modeling that make it seem as though a known personality is saying or doing something that they did not say or do. Deepfakes are used to create credible messaging that is false.

Although some kinds of misinformation can be harmless, many kinds of disinformation can be extremely harmful to individuals and societies. For example, Covid-19 anti-vaccination disinformation on social media in 2021 led to vaccination hesitancy in many countries, which led to greater spread of the disease and death. Other disinformation has political motives that are meant to destabilize a nation.

18.2.3 Hate Speech and Inciting Violence

Whether false or true (disinformation or not), hate speech (abusive language against a particular group) attracts attention. Traditional media typically does not disseminate hate speech. The terms and conditions of many social media platforms also do not allow for hate speech and provide mechanisms for users to flag it. Nevertheless, since the problem of defining and moderating hate speech at the scale of worldwide digital platforms is difficult, much hate speech does get posted in social media platforms and then amplified via information filtering algorithms because it is so engaging.

---

[8]Soroush Vosoughi, Deb Roy, and Sinan Aral. "The Spread of True and Fake News Online." In: *Science* 359.6380 (Mar. 2018), pp. 1146–1151.

[9]Karen Hao. "How Facebook and Google Fund Global Misinformation." In: *MIT Technology Review*. URL: https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait, 2021.

Messages on social media platforms and actions in the real world are closely intertwined.[10] Hate speech, offensive speech, and messages inciting violence on digital platforms foment many harms in the physical world. Several recent instances of hateful violence, such as against the Rohingya minority in Myanmar in 2018 and the United States Capitol Building in 2021, have been traced back to social media.

18.2.4 Alternatives

You've seen how maximizing engagement leads to negative externalities in the form of real-world harms. But are there proven alternatives you could use in the machine learning algorithm running Upwe.com's information retrieval system instead? Partly because there are few incentives to work on the problem among researchers within 'big tech,' and because researchers elsewhere do not have the ability to try out or implement any ideas that they may have, the development of alternatives has been few and far between.[11]

Nevertheless, as you develop the paradigm for Upwe.com, the following are a few concepts that you may include. You may want the platform to maximize the truth of the factual information that the user receives. You may want the platform to always return content from a diversity of perspectives and expose users to new relations with which they may form a diverse social network.[12] You may wish to maximize some longer-term enjoyment for the user that they themselves might not realize is appropriate for them at the moment; this paradigm is known as *extrapolated volition*. Such concepts may be pursued as pre-processing, during model training, or as post-processing, but they would be limited to only those that you yourself came up with.[13] A participatory value alignment process that includes members of marginalized groups would be even better to come up with all of the concepts you should include in Upwe.com's paradigm.

Furthermore, you need to have transparency in the paradigm you adopt so that all members of society can understand it. Facts and factsheets (covered in Chapter 13) are useful for presenting the lower-level test results of individual machine learning models, but not so much for institutional trust (except as a means for trained auditors to certify a system). CP-nets (covered in Chapter 14) are understandable representations of values, but do not reach all the way back to the value system or paradigm. It is unclear how to document and report the paradigm itself, and is a topic you should experiment with as you work on Upwe.com.

---

[10]Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. "The Effect of Extremist Violence on Hateful Speech Online." In: *Proceedings of the AAAI International Conference on Web and Social Media*. Stanford, California, USA, Jun. 2018, pp. 221–230.

[11]Ivan Vendrov and Jeremy Nixon. "Aligning Recommender Systems as Cause Area." In: *Effective Altruism Forum*. May 2019.

[12]Jianshan Sun, Jian Song, Yuanchun Jiang, Yezheng Liu, and Jun Li. "Prick the Filter Bubble: A Novel Cross Domain Recommendation Model with Adaptive Diversity Regularization." In: *Electronic Markets* (Jul. 2021).

[13]Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. "What Are You Optimizing For? Aligning Recommender Systems with Human Values." In: *Proceedings of the ICML Participatory Approaches to Machine Learning Workshop*. Jul. 2020.

18.3 Taxes and Regulations

There are few incentives for existing, entrenched platforms to pursue paradigms different from engagement maximization in the capitalist world we live in. Upwe.com will find it very difficult to break in without other changes. Short of completely upending society to be more relational via structures such as village-level democracy and self-reliance promoted by Mahatma Gandhi or anarchism,[14] the primary ways to control the harms of maximizing engagement are through government-imposed taxes and regulation spurred by a change in societal norms.[15] The norms should value the wellbeing of all people above all else. Viewing machine learning for information filtering as an institution rather than an individual, it is not surprising that the people who support interventions for controlling the negative externalities of the systems are those who have strong trust in institutions.[16] Society may already be on a path to demanding greater control of digital media platforms.[17]

While building up and developing Upwe.com, you should take a page out of Henry Heinz's playbook (remember from the preface that in addition to developing trustworthy processed food products, he lobbied for the passage of the Pure Food and Drug Act) and push for stronger regulations. Some possible regulations recommended by the Aspen Institute are:[18]

1. *High reach content disclosure*. Companies must regularly report on the content, source, and reach of pieces of knowledge that receive high engagement on their platform.
2. *Content moderation disclosure*. Companies must report the content moderation policies of their platform and provide examples of moderated content to qualified individuals.
3. *Ad transparency*. Companies must regularly report key information about every ad that appears on their platform.
4. *Superspreader accountability*. People who spread disinformation that leads to real-world negative consequences are penalized.
5. *Communications decency control on ads and recommendation systems*. Make companies liable for hateful content that spreads on their platform due to the information filtering algorithm, even if it is an ad.

Many of these recommended regulations enforce transparency since it is a good way of building institutional trust. However, they do not provide governance on the paradigm underlying the platform because it is difficult to measure the paradigm. Nevertheless, they will control the paradigm to some extent. If social media platforms are deemed *public utilities* or *common carriers*, like telephone and electricity providers, then even more strict regulations are possible. Importantly, if you have designed

---

[14]Brian Martin. *Nonviolence versus Capitalism*. London, England, UK: War Resisters' International, 2001.
[15]Daron Acemoglu. "AI's Future Doesn't Have to Be Dystopian." In: *Boston Review*. URL: https://bostonreview.net/forum/ais-future-doesnt-have-to-be-dystopian/, 2021.
[16]Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. "Misinformation Interventions are Common, Divisive, and Poorly Understood." In: *Harvard Kennedy School Misinformation Review* 2.5 (Sep. 2021).
[17]Throughout the chapter, the governance of platforms is centered on the needs of the general public, but the needs of legitimate content creators are just as important. See: Li Jin and Katie Parrott. "Legitimacy Lost: How Creator Platforms Are Eroding Their Most Important Resource." URL: https://every.to/means-of-creation/legitimacy-lost, 2021.
[18]Katie Couric, Chris Krebs, and Rashad Robinson. *Aspen Digital Commission on Information Disorder Final Report*. Nov. 2021.

Upwe.com to already be on the right side of regulations when they become binding, you will have a leg up on other platforms and might have a chance of being sustainable.

In parallel, you should also try to push for direct ways of controlling the paradigm rather than controlling the negative externalities because doing so will be more powerful. Regulations are one recognized way of limiting negative externalities; Pigouvian taxes are the other main method recognized by economists. A *Pigouvian tax* is precisely a tax on a negative externality to discourage the behaviors that lead to it. A prominent example is a tax on carbon emissions levied on companies that pollute the air. In the context of social media platforms, the tax would be on every ad that was delivered based on a targeting model driven by machine learning.[19] Such a tax would directly push 'big tech' companies to change their paradigm while leaving the Upwe.com paradigm alone.

Seeing out your vision of an Upwe.com that contributes to the wellbeing of all members of society may seem like an insurmountable challenge, but do not lose hope. Societal norms are starting to push for what you want to build, and that is the key.

18.4 Conclusion

- There is so much and such complicated knowledge in our world today that it is impossible for anyone to understand it all, or even to verify it. We all have epistemic dependence on others.
- Much of that dependence is satisfied by content on the internet that comes to us on information platforms filtered by machine learning algorithms. The paradigm driving those algorithms is maximizing the engagement of the user on the platform.
- The engagement maximization paradigm inherently leads to side effects such as filter bubbles, disinformation, and hate speech, which have real-world negative consequences.
- The machine learning models supporting content recommendation on the platforms is so disconnected from the experiences of the general public that it does not make sense to focus on models' interpersonal trustworthiness, which has been the definition of trustworthiness throughout the book. An alternative notion of institutional trustworthiness is required.
- Institutional trustworthiness is based on governance mechanisms and their transparency, which can be required by government regulations if there is enough societal pressure for them. Transparency may help change the underlying paradigm, but taxes may be a stronger direct push.
- A new paradigm based on relational ethics is needed, which centers truth, a diversity of perspectives, and wellbeing for all.

> "I nod to you and up we come."
>
> —Norbert Buskey, band teacher at Fayetteville-Manlius High School

[19]Paul Romer. "A Tax To Fix Big Tech." In: *New York Times* (7 May 2019), p. 23.