Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Article 6

2022

Online Extremism, AI, and (Human) Content Moderation

Michael Randall Barnes
Australian National University
michael.barnes@anu.edu.au

Recommended Citation
Barnes, Michael Randall. 2022. "Online Extremism, AI, and (Human) Content Moderation." *Feminist Philosophy Quarterly* 8 (3/4). Article 6.

Online Extremism, AI, and (Human) Content Moderation

Michael Randall Barnes

Abstract

This paper has three main goals: (1) to clarify the role of artificial intelligence (AI)—along with algorithms more broadly—in online radicalization that results in "real world violence," (2) to argue that technological solutions (like better AI) are inadequate proposals for this problem given both technical and social reasons, and (3) to demonstrate that platform companies' (e.g., Meta, Google) statements of preference for technological solutions functions as a type of propaganda that serves to erase the work of the thousands of human content moderators and to conceal the harms they experience. I argue that the proper assessment of these important, related issues must be free of the obfuscation that the "better AI" proposal generates. For this reason, I describe the AI-centric solutions favoured by major platform companies as a type of obfuscating and dehumanizing propaganda.

Keywords: artificial intelligence, social media, content moderation, online extremism, radicalization, propaganda

Introduction

This paper addresses the role that artificial intelligence (AI) plays in online radicalization, along with its potential role in combatting online extremism.[1] I argue there is a plausible case to be made that AI is partly responsible for some harmful events, and that the hope that improvements in AI might correct for these harms is overstated, at best. Moreover, I argue that the rhetoric around AI from some of its biggest champions serves an ideological and propagandistic function that should be uncovered and resisted. In short, the idea that "better AI" will be our salvation is a distraction from the severity of the problems for which large the technology companies (Big Tech) bear some responsibility. Those problems deserve our direct attention, and "techno-solutionism" offers an undeserved shield to powerful actors.

[1] A better term for my topic is *automated decision systems*, though even that is imprecise. In any case, I aim to include AI and algorithms more broadly in my scope but choose to use the language of AI partly to mimic the rhetoric of Big Tech.

My focus here is on cases like the horrific mass murders at mosques in Québec City in 2017 and in Christchurch in 2019 (and many other examples), which demonstrate the importance of addressing the worrying trend of online hate spilling into the "real" world. We are quickly coming to understand that "self-radicalization," as in the cases of Dylann Roof in Charleston and Alek Minassian in Toronto, is a growing problem that demands our attention. But we are only beginning to understand how violent hate spreads on our new communication channels. Emerging research—along with confessions from the radicalized themselves—suggests that online platforms play a vital role in accelerating the rise of extremism at both the individual and the societal levels. By leading users down a path of radicalization and serving as an essential medium connecting formerly isolated extremists, disparate platforms like Facebook, YouTube, Gab, Discord, 8kun, and more are implicated in this problem. And after years of public pressure from frustrated academics, victims' groups, and civil rights organizations, many of these platforms have switched their responses to these charges from overall dismissiveness to pledges to do better.

A major concern, though, is precisely how these technology companies have chosen to address this problem. Some initial reflection reveals a few complications to these issues from the platforms' perspective. The motivation to keep users online makes platforms hesitant to remove hateful content even when it clearly violates their terms of service agreements—including content from political candidates and heads of state, whose visibility is much greater than most users. This resistance is partly explained by the desire of these private companies to appear politically neutral, as accusations of bias undermine their overall efforts for growth. But platform companies are also aware that too much vile content is itself a great threat to their business, and they work hard to keep "objectionable" content like pornography and graphic violence out of users' feeds. This pair of pressures leads many platform companies to prefer "black-box" algorithmic solutions for moderation decisions, as these come with a veneer of objectivity. But research consistently reveals that this technology is not unbiased.[2] Rather than eliminating human bias, technological solutions often replicate and amplify prejudices. Still, CEOs like Meta's Mark Zuckerberg have made it clear that their main efforts to clean up their platforms will be based on algorithmic (and, increasingly, AI-based) solutions. For those who have been watching, there is a sense that this is all too little, too late. Hate speech and misinformation run amok online, recommendation algorithms still lead users to hate groups, and many critics believe that Big Tech *still* does not appreciate the severity of the problem on their hands.

Furthermore, the idea that this is mainly a *technological* problem, and so one that calls for a technological solution, deserves scrutiny. Consider moderation: While

[2] See, for example, Noble (2018), Benjamin (2019), O'Neil (2016), and Eubanks (2018).

algorithmic moderation has the potential to limit audience exposure to extremist content, it is not the panacea it is occasionally presented as—usually by tech companies themselves. Algorithmic moderation is generally ineffective at altering the habits of the already radicalized, is prone to false positives/negatives, and is sometimes easy to game. Put simply, it just does not live up to the hype. Fully understanding why this approach is ultimately ineffective, though, requires seeing the broader social and human aspects of these issues.[3]

A compelling entry point for understanding these issues, I believe, rests in the under-told story of the thousands of *human* content moderators employed by companies who profit off user-generated content. As things stand, Big Tech relies on these moderators—often employed via third-party mediators—to sift through countless toxic posts every day, resulting in a mix of psychological harms. Most of these moderators are underpaid contactors, often (though not always) located overseas, and are not the image of a happy, spoiled employee these companies usually project. Yet they are essential workers in the global information supply chain—and a full account of online hate must explore the harms they experience. And, as I will argue, the neglect they are shown in the public stances of Big Tech reveals an overall ideology in which technological "progress" is valued over human flourishing to a troubling degree. In short, proposals that focus on the future *potential* of algorithmic solutions dehumanize these workers, obfuscate the harms they face, and complicate the debate about the distribution of responsibility in actually addressing these challenges. For this reason, I describe these AI-centric solutions as a type of propaganda.

I begin in section 1 by providing an overview of how the proprietary algorithms of Big Tech may be said to promote radicalization and extremism online. YouTube and Facebook will be my primary examples here, as these have received the most attention from critics. Then, in section 2, I highlight one of the main the proposals put forward by these Big Tech firms: that better AI will be the primary tool used to overcome these problems. I then critically assess this proposed solution, arguing that it is, at best, only a partial solution that is inadequate to address current problems, and more importantly, one that obscures the discussion around these issues. In section 3, by considering the situation of content moderators, I consider how the preference for *technological* solutions serves an ideological function for Big Tech and its champions. That is, its primary purpose—whether intended as such or not—is as propaganda that distracts us from the real, avoidable, human harm companies like

[3] For a revealing overview of the complexities of content moderation, see Gillespie (2018). Similar considerations also apply to other applications of AI outside the context of the moderation debate. See Zimmermann, Di Rosa, and Kim (2020).

Meta Platforms (who own Facebook) and Google (who own YouTube) contribute to worldwide.

1. Technology and Radicalization

Not long before he went on to murder eleven people at Pittsburgh's Tree of Life synagogue, Robert Gregory Bowers posted to the social media site Gab. He wrote, "HIAS [the Hebrew Immigrant Aid Society] likes to bring invaders in that kill our people. I can't sit by and watch my people get slaughtered. Screw your optics, I'm going in" (Gessen 2018). This was one of a number of mass shootings that were preceded by posts on social media—sometimes manifestos, other times simply declarations of intent—or, in the more horrific cases, were live-streamed. The relationship between mass shootings (often targeting Jews, Muslims, and/or immigrant communities) and social media has gone mainstream. As *New York Times* columnist Charlie Warzel (2019) writes, "Mass shootings have become a sickening meme," adding that "it's becoming increasingly difficult to ignore how online hatred and message board screeds are bleeding into the physical world—and how social platforms can act as an accelerant for terroristic behavior."

This increased visibility has led to an increase in critical scrutiny. The notorious imageboard 8kun (formerly 8chan) struggled to stay online as mainstream culture became aware of its apparent role in radicalizing mass murderers. Social networking sites Gab and Parler have also had difficulties maintaining operations amidst the bad press that follows mass shootings and would-be insurrections. These sites are all comparatively much smaller, and their reach less widespread, than the big platforms like Facebook and YouTube. But just because these bigger sites don't explicitly cater to extremists doesn't mean they aren't also implicated.

To be sure, extremist violence is not new, but the online roots of many recent mass shootings force us to consider to what extent contingent socio-technological developments (e.g., social media) have increased or exacerbated long-standing problems. There are at least two promising avenues with which we can justify this inchoate claim. First, we can note that, much like previous revolutionary developments in communications technologies, social media brings much larger numbers of people together than was previously possible. As such, formerly isolated individuals are now able to find others who share their beliefs and ideals. When it is violent extremists who connect with like-minded fellows, an increase in real-world violence is the (likely) result. That is, social media brings potential perpetrators together, and this connection plays a nonnegligible role in bringing about more violence. I'll call this avenue the *social route*.

The second, distinct but related way to justify my claim of a causal connection between online platforms and increased real-world violence is to explain how radicalization may occur when an individual is served more and more "extreme"

content, culminating in their adoption of an ideology that encourages violent acts. This is what is said to happen when an individual falls down a "rabbit hole" online, leading one from questions about the fairness of affirmative action, say, all the way to support for white supremacy. A key component of this story that emphasizes the platform's role in this process is the notion that these steps towards radicalization are laid out by recommendation algorithms employed by YouTube and Facebook. I'll call this avenue the *individual route*. I clarify and develop these two routes below.

1.1. The Social Route

At the heart of the social route is the notion that connecting previously isolated extremists generates new potential for violent acts. That social media brings people together is uncontroversial. Indeed, until recently, Facebook itself stated that its core mission is to "give people the power to build community and bring the world closer together."[4] But the more specific claim that social media connects *extremists* requires elaboration. Unfortunately, it is evident that this is the case.

There is, in fact, a long history of extremists being early adopters of networked technologies, often with the explicit goal of both increasing connection among existing members and recruiting new ones. For the first few decades of the internet, this mainly consisted in hosting their own websites and forums whose existence circulated through word of mouth. Connection beyond the existing community remained relatively difficult. The rise of social media changed this. As Laura Smith (2021) writes:

> On platforms like Twitter and Facebook, extremists could organize and share information, often in plain sight. Instead of thousands of people reading online bulletin boards, tens of millions were seeing racist Pepe the Frog memes, "white genocide" rhetoric and conspiracy theories about Democrats running child trafficking rings.

And as examples like the 2017 "Unite the Right" rally in Charlottesville, Virginia, show, violent groups will use the online tools available to them to organize. The problem is even more acute in regions of the world where internet access is disproportionately mediated through a single platform like Facebook—places where the company sought growth without first sufficiently understanding existing regional conflicts and failed to

[4] This has changed a bit over the years, but the general idea has remained constant. See https://www.facebook.com/facebook/about/ for the current version, and see https://web.archive.org/web/20201206131854/https://www.facebook.com/pg/facebook/about/ for an archived version from December 6, 2020 that contains the old motto.

invest in moderators who understood the local languages. The experience of Facebook in Myanmar is a particularly grim case.[5]

That problems like these would emerge on social media is, in a sense, inevitable. As Tarleton Gillespie (2020, 1) puts it, "As social media platforms have grown, so has the problem of moderating them." When the userbase is small, problems can often be addressed as they emerge. But that is no longer the case: the userbase is much larger and the problems are much more difficult to address. And seeing why these problems are more complicated clarifies why extending blame to social media companies themselves is apt. Beyond the sheer difficulty posed by the tremendous increase in scale, one source of our current problems is that users are now "linked less by the bonds of community and more by recommendation algorithms and social graphs" (Gillespie 2020, 1). In other words, platforms are *pushing* people together who might otherwise remain disconnected. In this way—and others to be elaborated below—platforms like Facebook are contributing a positive act—namely, they are offering their users *invitations* to connect. This shifts their role from (an arguably) neutral platform to one where they actively shape and contribute to the further acts their users perform.

Here we can move from the abstract claim that, say, Facebook is responsible for a rise in extremist violence to the more concrete claim that Facebook's *algorithms*—many of which deploy innovative techniques of machine learning (ML) and/or AI—are responsible. While the data to support this claim is not all publicly available, largely held as it is by notoriously untransparent companies, some brief examples demonstrate the basics of the accusation.

Facebook has long made "friend" suggestions to "connect" users to one another, and the company has often done this by mining data in arguably nefarious ways. They do so under the guise of "building community." The more recent push within the platform to promote Facebook Groups to the centre of the user experience has increased the role of their recommendation algorithms, from connecting users who shared friends or workplaces, to connecting users according to interests. The themes of these groups, however, include interests as disparate as knitting, anime, and the violent overthrow of democratic governments, as well as various conspiracy theories.

For example, the growth of the QAnon conspiracy theory across social media was, for those who witnessed its early moves, a strange sight to behold. In a relatively short time, QAnon went from a fairly small phenomenon localized to obscure parts of the internet to mainstream news, and it was at the forefront of the violent riot that

[5] See Stecklow (2018). And here it is useful to remember that another one of Facebook's mottos was "Move fast and break things." See Statt (2014) for an account of the motto's place at Facebook and why it was repealed.

broke into the US Capitol. This growth would not have been possible without social media, and more specifically without the AI-driven recommendation engines that pushed QAnon across the internet. As Renée DiResta (2020) describes it:

> Algorithmic recommendation engines accelerated [QAnon groups'] growth and cross-pollinated their beliefs. Over time, these engines nudged anti-vaxxers and flat-earthers to join QAnon groups and pushed QAnon videos to far-right political communities.

The growth of QAnon groups persisted alongside assertions by Facebook that groups that violated its community standards policies would be banned. Similarly, antigovernment militia groups continued to be recommended to Facebook users even after Facebook claimed they were cracking down on these groups and would stop recommending them. This happened because algorithms trained to optimize engagement led (and continue to lead) users to these groups, and these recommendations are automated, like so much else users see on the platform.[6]

What this demonstrates is that the combination of technology—AI-driven recommendation engines trained to increase engagement above all else—and company policy—lack of robust enforcement of existing community standards—plausibly played a role in swelling the ranks of violent conspiracy groups and militias. While definitive proof is not available, it does seem as though the January 6, 2021, "Save America" rally was planned in large part on Facebook, in closed groups. What would have happened had Facebook not trained its AI in a way that led users towards conspiracy and militia groups is, of course, impossible to know. But Facebook's actions after the fact—deflecting blame to other, much smaller platforms; pausing recommendations on political groups; removing QAnon groups with more determination—suggest that they believe these actions are still worth taking now.

At the same time, it's worth remembering that while AI seems to play a role in exacerbating extremism at the social level, it is in no way solely responsible. This is true even when our eyes are trained on social media platforms. Both WhatsApp and Telegram have been implicated in extremist organizing and mob violence, but neither relies on AI or recommendation algorithms to push users together. Users, all on their own, are capable of finding one another and entrenching themselves in violent rhetoric. With that in mind, I'll now turn to the second route whereby technology may be claimed to increase radicalization.

[6] Only recently have social media companies begun to actually crackdown on QAnon content. Early reports suggest their efforts have been successful, though only time will tell if it is a mere band aid; see Dwoskin and Timberg (2021)

1.2. The Individual Route

I argued above that platforms plausibly increase extremism by connecting previously isolated radical individuals. But an additional problem is that the platforms serve an important function in radicalizing previously nonradicalized individuals. At first blush, this may seem to only be a problem for the especially toxic sites. While curiosity might first push a user towards a site like 8kun, the dynamics of that environment—where escalations in toxicity are explicitly encouraged and users are inundated with nonstop offensive content—could play a radicalizing role that has no counterpart on the relatively tamer mainstream platforms.

However, platforms like Facebook and YouTube do indeed seem to be implicated in the problem of radicalizing individuals. As such, this is not merely a problem of the "fringe" internet, but something much more central to our current social-media-saturated landscape. Moreover, with these mainstream platforms, technologies like AI are also implicated in this problem.

The case is perhaps strongest for YouTube, whose recommendation algorithm has received significant criticism for pushing users towards more "extreme" content.[7] Journalists and academics have documented how YouTube's default setting of autoplaying "related" videos once a selected video ends can lead viewers to conspiracy theories and generally more ideologically radical content than what is specifically selected. That the recommendation engine might push users in a certain direction matters because, as YouTube stated in 2018, more than 70 percent of the time users spend watching videos on the platform is driven by its recommendations (Solsman 2018). Given the incredible scale of YouTube, which reaches billions of users worldwide, the possibility that this engine is autoplaying potentially radicalizing content is concerning. And available evidence appears to back this up.[8]

The general thrust of the problem is that in optimizing for more viewership hours—because the longer viewers are on the platform, the more advertisements can be served—YouTube's recommendation system can take someone from innocent search terms about music or dating to videos showcasing "pickup artists," and from there to videos promoting (toxic) masculinity and white nationalism, as well as a litany of conspiracy theories. This can occur, moreover, with very little action on the part of the viewer aside from simply sitting back and consuming content. The main driver on this pathway, in other words, is the AI-driven recommendation algorithms YouTube employs.

Alfano et al. (2021) call this process "technological seduction," and with regards to YouTube, it occurs via "technologically-mediated cognitive pressures and

[7] See, for example, Lewis (2018, 2020), Zadrozny (2021), and McCrosky and Geurkink (2021).

[8] See, for example, Alfano et al. (2021).

nudges that subtly but systematically induce acceptance of problematic beliefs" (Alfano et al. 2021, 837).[9] For YouTube, the nudges are fairly blunt, as they don't simply suggest a follow-up video but indeed play that video absent any action on the viewer's part. It is, in other words, the default course on the platform. Assuming this leads to "more extreme" content, these nudges have the effect of leading users down a rabbit hole of conspiracy content that may shape their beliefs and actions outside of the platform.[10] This is therefore another way social media companies contribute to this, admittedly complex, problem.

Something similar occurs with Facebook's Newsfeed, where the algorithm dictates what content appears on user's screens. Here, though, the content is often in the form of posts written by other users—though the presence of paid/promoted material complicates things. Other user actions like Shares, Likes, and other Reactions play some role here too, but these are best seen as simply another factor among many that act as input to the algorithms that ultimately decide what users see. These algorithms act as amplifiers of certain pieces of content and demoters of others. In some cases, these algorithms reward inflammatory, "borderline" content because they receive more user engagement. In other words, Facebook's algorithms amplify content that nearly-but-not-quite violates its own community standards and puts this content into more users' feeds (Hao 2021a).

The specifics of algorithmic amplification are complicated by both the complexity of the algorithms involved, as well as their opacity. And yet, as Michaelson, Pepp, and Sterken (2021) observe, "amplification has taken on a structural significance in online speech that is totally unprecedented in earlier, offline speech." They distinguish amplification from "appropriation" and note that "in cases of amplification, in contrast to cases of appropriation, the amplifier takes reasonable steps to see to it that the original speaker retains credit for their contribution to the conversation" (ibid.). That is, "unlike offline means of amplification," Michaelson, Pepp, and Sterken write, "online amplification via retweeting on Twitter or sharing on Facebook is designed to *automatically credit the content to its original creator*, as a share or retweet preserves something like a copy of the original post along with its metadata" (ibid.; emphasis added). However, while this effort to retain credit may be laudable for individual users, it has a very different impact when we consider how the

[9] Alfano, Carter, and Cheong (2018) distinguish between "top-down" and "bottom-up" technological seduction. The case which concerns us here is one of bottom-up, "which occurs as the result of technological systems creating suggestions based on aggregated user data" (Alfano et al. 2021, 838).

[10] At the same time, it's important that we do not scapegoat the YouTube algorithm as the sole mechanism through which YouTube encourages radicalization; see Lewis (2018). And for the polarizing potential of social media itself, see Lynch (2016).

most important amplifiers on social media (i.e., the platform's algorithms) use this feature to hide their own actions. This design feature not only credits the original creator but has the effect of emphasizing the original post while downplaying the novel act of reproducing it for more to see. That is, in algorithmic amplification, a user's words are taken from their original context and inserted in many new ones, increasing their reach, while at the same time this reproduction—including the fact that it is, as a type of repetition, its own (speech) act—is played down. But this act of amplification, and the decisions that prioritize certain types of content over others, again renders platforms like Facebook and their algorithms culpable.

Of course, the effects here shouldn't be overstated. It's rare that someone will simply accept the outlandish tenets of the QAnon conspiracy simply because they were led to a few videos online or read a few inflammatory posts. But the effect is not negligible either. Both Facebook and YouTube have over two billion monthly active users, and a mere tenth of a percent of that figure is two million people. So, the potential to radicalize even a tiny fraction of their user base is still a problem that platform companies—and broader society—must take seriously. For users who are isolated—either because of the COVID-19 pandemic or unrelated causes—and perhaps vulnerable in other socially relevant ways, encountering this type of content can be a source of reassurance and an antidote to broader feelings of powerlessness. Platforms arguably hold a special duty of care towards these users, for whom this content poses particular dangers as a result of social isolation, mental health concerns, or other relevant factors. The need for platform companies to take this problem seriously is further underscored by the fact that the ultimate harms of extremist violence typically fall on the already-marginalized, whose well-being and safety concerns plausibly deserve priority. Therefore, even if the paths I've presented are a miniscule part of the average experience of most social media users, the problems they present are still urgent when considered in absolute terms. The decisions platform companies make and the actions they take deserve scrutiny, as they open these paths up to users and are culpable—if only partly—for the harms they lead to.

1.3. Convergence

In addition to offering a (in some senses) comforting worldview, extremist content also often offers a community, and this is where the social and individual routes converge. Rarely do violent extremists self-radicalize entirely in isolation, absent any dialogical contact with others. A further factor in these narratives lies in places like the comment sections of platforms like YouTube, where users can engage with one another. It's also common for the creators of these videos to cross-reference and host one another, encourage viewers to follow them across social media, and add

to the discussion in comments. These activities all play a role in fostering a sense of community across the internet.

And while total self-radicalization is rare, it does seem to occur, as cases like Dylann Roof's and Alexandre Bissonnette's appear to demonstrate. But even in these cases, it's incorrect to claim these individuals weren't part of any communities. That is, we should note that the type of community at issue is not necessarily one whose members engage in frequent dialogical exchange. Simply being an adherent to the QAnon conspiracy can place one in that community even without tight social bonds. Indeed, this sort of silent community is a regular fixture in extremist groups, and the long history of white nationalist groups offers a compelling case to learn from. Because of a mixture of actual and perceived persecution, white nationalists have (or had) tended to keep most communications in private spaces, fearful of government or antiracist surveillance. While this decreases the capacity for the type of constant encouragement towards violence common on 8kun, it has a sinister correlate: advocacy for "leaderless resistance."

Within extremist movements where leaderless resistance is encouraged, a recurring refrain is the notion that members of the movement cannot and should not wait for direct orders. Rather, they should take it upon themselves to plan and execute violent attacks where they are able. In this way, the social and individual routes converge once again. Once an individual sees themselves as part of a community that advocates for leaderless resistance, they may no longer require the actual encouragement of the community to feel its pull. When shooter Robert Gregory Bowers wrote, "Screw your optics, I'm going in," he was referring to the intercommunity debate within white nationalist groups about the best strategy to enact their utopia. His urge to act was both a response to the community he was a part of and something more individualistic.

Again, in none of these cases is technology solely responsible. There is no example of AI coercing an individual to watch (and believe) certain conspiracy videos, nor do we see recommendation algorithms forcing Facebook users to join militia groups. But there is encouragement, and occasionally seduction, often in the form of invitations, recommendations, and amplifications. These encouragements, in combination with other relevant factors, plausibly play some role in increasing extremism within our progressively more connected society. The fact that the Big Tech companies themselves have taken pains to address these issues offers compelling evidence that there is something here. In the following sections I turn to the strategies Big Tech has developed in response to these big challenges in order to argue that one strategy—namely, its focus on improving its AI—is an inadequate response, and the attention it receives can be considered a form of propaganda that renders discussion on this topic more difficult and hinders the proper allocation of responsibility.

2. The "Better AI" Approach of Big Tech

The previous section served to add some detail to the claim that platform companies and the technologies they employ—especially AI and other algorithms—contribute to a rise in violent extremism. In this section I will address how the companies themselves have responded to these issues, focusing on their claim that the solution lies in AI itself. After doing so, I will begin to criticize this approach, focusing first on more technical limitations before moving on to its human costs in section 3.

The most explicit champion of AI's (future) capabilities is possibly Mark Zuckerberg. In his April 2018 congressional testimony, the Meta Platforms (then known as Facebook) CEO invoked the potential of AI to someday solve the problems for which he was then being frequently criticized.[11] And to be sure, Meta/Facebook is investing in the development of this technology and has seen some success. After receiving international criticism for the platform's role in ethnic violence in Myanmar, Facebook "improved its Myanmar-language hate-speech classifiers, leading to a 39% increase in takedowns from automated flags in only six months" (Gorwa, Binns, and Katzenbach 2020, 2). And, after the 2019 Christchurch terrorist streamed his shooting rampage on Facebook Live, the company worked to aggressively remove the many re-uploads that occurred in the aftermath. According to Facebook, "In the first 24 hours, versions of the video had been uploaded at least 1.5 million times, and some 80% of those videos, around 1.2 million, were blocked automatically before they could be uploaded" (Gorwa, Binns, and Katzenbach 2020, 1–2).

The massive size of Facebook and other platforms is a major reason these companies are turning to automated systems to address unwanted content. The self-moderated spaces of the early internet have largely been left behind, and platforms have instead turned to "commercial content moderation" (Roberts 2019), with dedicated teams and systems operating behind the scenes and at scale. And as impressive as the massive removal of the Christchurch videos was, it also helps to demonstrate some of the limitations of the "improved AI" approach.

In the hour or so before Facebook could remove the original stream, the Christchurch video had already been downloaded. It would later be reposted, often with slight alterations that made it difficult for automated systems to detect. While simple copies were removed en masse, usually before they could ever reach users, many altered copies slipped through. In this lies a basic lesson about algorithmic moderation: it can achieve massive results when applied to known problems, but

[11] As the *New York Times* reported it: "Before Congress last year, Mr. Zuckerberg testified that Facebook was developing machine-based systems to 'identify certain classes of bad activity' and declared that 'over a five- to 10-year period, we will have A.I. tools' that can detect and remove hate speech" (Metz and Isaac 2019).

deviations and innovations in content pose difficult challenges. This is particularly true when we distinguish between two common types of algorithmic moderation systems that have different capabilities.

Gorwa, Binns, and Katzenbach (2020, 3) explain this well. At a general level, they write, algorithmic content moderation "involves a range of techniques . . . [that] aim to identify, match, predict, or classify some piece of content (e.g., text, audio, image or video) on the basis of its exact properties or general features." However, they add:

> One major distinction can be made between systems that aim to *match* content ("is this file depicting the same image as that file?"), and those that aim to *classify* or *predict* content as belonging to one of several categories ("is this file spam? Is this text hate speech?"). (Gorwa, Binns, and Katzenbach 2020, 3; emphasis added)

Matching systems, they write,

> Typically involve "hashing," i.e., the process of transforming a known example of a piece of content into a "hash"—a string of data meant to uniquely identify the underlying content. Hashes are useful because they are easy to compute, and typically smaller in size than the underlying content, so it is easy to compare any given hash against a large table of existing hashes to see if it matches any of them. (4)

It is this type of system that was responsible for removing most copies of the Christchurch video.[12] Facebook is a member of the Global Internet Forum to Counter Terrorism (GIFCT), a group created by Facebook, Google, Twitter, and Microsoft and committed to industry collaboration in efforts to combat illegal online hate speech. By sharing best practices, these four platform companies develop automated systems and share a "hash database" of terrorist content, isolating their digital fingerprints. In a matter of days, Facebook had uploaded "800 visually-distinct videos related to the attack" (Sonderby 2019).

[12] To be precise, Gorwa, Binns, and Katzenbach (2020, 4) point out that a related but novel technique, "perceptual hashing," is the "most suitable and robust for content moderation" as it enables matching for content that has had some irrelevant (to the human eye) content altered, and it is likely this type of system Facebook and the GIFCT employed in removing videos of the Christchurch terrorist attack—though the specific details are not public.

The other broad category of algorithmic moderation is *classification* systems. In this case, a system "assesses newly uploaded content that has no corresponding previous version in a database; rather, the aim is to put new content into one of a number of categories" (Gorwa, Binns, and Katzenbach 2020, 5). The categories that interest us here are items like "hate speech," "offense," "harassment," "terrorism," and so on, but these systems can also be used to detect features like "nudity," "sexual content," and much else. The classification tools used in content moderation often involve machine learning—that is, "the automatic induction of statistical patterns from data" (5)—whose capabilities have greatly increased in recent years, alongside developments in Natural Language Processing (NLP) and computer vision.

Both systems, however, are imperfect. Matching is useful for identifying a new piece of content against an existing database to see if it's a copy. This is essential in situations where many users are uploading a known piece of unwanted content, as in the Christchurch case, and platforms have used it to clamp down on terrorist content and child sexual abuse material (CSAM) at scale. But as Gillespie (2020, 3) puts it, "This is automation, but it is hardly AI." He adds that a consequence of this is that "recent claims by platforms of successful automated moderation are overstated" (3). By this he means to deflate some of the enthusiasm implicit in Facebook's boasts about hate-speech content being flagged by Facebook's systems *before* users reported it. As he explains:

> At least for now, the overwhelming majority of what is being automatically identified are copies of content that have already been reviewed by a human moderator. Stats like these are deliberately misleading, implying that machine learning (ML) techniques are accurately spotting new instances of abhorrent content, not just variants of old ones. (Gillespie 2020, 3)

And this obfuscation matters because classification systems are also limited. As these systems require massive data sets to train on, there is an inevitable lag in their ability to autonomously label new content. But hate speech is a constantly evolving category. New groups become targets, and new terms and phrases are used to dehumanize. As Hao (2021a) reports, "An algorithm that has learned to recognize Holocaust denial can't immediately spot, say, Rohingya genocide denial. It must be trained on thousands, often even millions, of examples of a new type of content before learning to filter it out." In general, AI is built on what already exists, but innovations in wording, phrasing, targets, tactics, and much else are always on the horizon, limiting its capabilities. AI can often be outwitted with a little experimentation and determination. And something that is effectively illegible to the AI may still be obvious, and harmful, to a human.

On the other hand, tools that over-block—that is, mistakenly remove inoffensive content—can be equally troublesome. Classification algorithms can have difficulty accounting for crucial factors such as context, sarcasm, and the fact that reporting an instance of hate speech is not itself an act of hate speech (in most cases). Journalists and regular users may want to share reports of hateful utterances or even terrorist propaganda but may find themselves at the mercy of a platform's automated moderation tools. Moreover, words that are used within a community often have different meanings when hurled at members of that community, but these are contextual features that are difficult for automated systems to handle satisfactorily. Indeed, some systems have been shown to have *racial dialect bias*, where they "systematically classify content aligned with the African American English (AAE) dialect as harmful at a higher rate than content aligned with White English (WE)" (Ball-Burack et al. 2021). This problem is amplified when we consider the global scale and constant quest for growth of Facebook and other platforms, and how they tend to treat the cultural meanings of local language users as afterthoughts, at best.

All this points to another frustrating aspect of algorithmic moderation: the affective impact it has on individual users. As Gillespie (2020, 3) puts it, the "gap between data-scale approaches and the individual experiences of them is slowly undermining the legitimacy of content moderation itself," and this reveals "a fundamental contradiction in using data-centric techniques for content moderation." By this, he means that

> moderation may be felt as an injustice—that the platform failed to understand the legitimacy of my post, failed to protect my right to speak. It may be a sense of absurdity, as when automated systems are wildly off the mark. It may be a feeling of suspicion, that these inscrutable methods only hide naked corporate self-interest. It may be a feeling of insignificance, when stilted, bureaucratic warnings offer a stark reminder that these megamachines have little concern for their users as individuals. (Gillespie 2020, 3)

Gorwa, Binns, and Katzenbach (2020, 3) echo this point when they note how large-scale algorithmic content moderation threatens to "decrease decisional transparency (making a famously non-transparent set of practices even more difficult to understand or audit)," among other worries.[13]

In sum, there are technical, social, and user-experience reasons to be suspicious about the "better AI" approach of Big Tech. Both matching and classification systems have serious limitations and are ill-equipped to deal with the

[13] See York (2021) for further issues on the user side of content moderation.

ever-evolving terrain of hate speech on their own. This all leads to a situation where regular users are unhappy, dedicated hate speakers can evade the algorithms, and attempts at counterspeech and documentation are thwarted. The hopes for and boasts about AI from Zuckerberg and others may be honest, as there is no shortage of optimists about AI's impending amazing capabilities. But their comments are also likely intended to appease lawmakers and investors, who have legitimate concerns that platforms are becoming too toxic. Furthermore, that Big Tech would go to the well of AI to address the issues brought on by AI is, despite appearances, not surprising. It is, as Geiger (2016, 791) puts it, representative of the "mindset prevalent in Silicon Valley, which sees these problems as technological ones requiring technological solutions." Still, the *potential* of AI is not a suitable response to the *current* problems of the platforms. This techno-solutionism takes a very narrow view of what it sees as the source of the problem and where intervention ought to occur.

This brings me to the most significant limitation of current AI systems—namely, that they rely on undervalued, invisible human labour. As Tubaro, Casilli, and Coville (2020, 7) note, "Only a small part of content moderation can be automated: any new types of data first require micro-workers to train future automated solutions." In addition to training, which involves labelling massive data sets and much else, the accuracy of these systems must be continually verified, and when accuracy cannot be achieved by the algorithms themselves, human workers often imitate the systems for the benefit of the unknowing user. There is, in other words, an entire world of labour that supports AI, and yet this aspect is rarely foregrounded by Big Tech.

With that in mind, I now turn to the underexplored human side of this equation, which I will argue is obscured by the AI-focused rhetoric of Big Tech. I argue that we should not simply see these comments as optimism about AI's capabilities, but when we put them alongside the phenomena this paper addresses—that is, extremism via algorithm, limits on algorithmic moderation, and the human cost of moderation—we can consider them a type of dehumanizing propaganda, as they erase the harms faced by thousands of current workers and lead the debate in an unhelpful direction.

3. Invisible Labour and Human Costs

While tech CEOs are more than willing to talk up their latest developments in harnessing AI and to grant media requests for interviews with the heads of their AI research teams, that openness is nowhere to be found with the topic of human content moderators.[14] Behind the scenes an army of employees play an indispensable

[14] In this section I draw heavily from Roberts (2019), as well as journalism and research including Newton (2019, 2020a, 2020b), Jones (2021a, 2021b), Solon (2018), and

role in sifting through the millions of posts made on social media each day.[15] It is their work that makes social networking platforms usable by keeping unwanted material from clogging our feeds. This usability, in turn, allows a company like Meta to become one of the most profitable companies in the world.

But despite its importance, content moderation is work that often occurs in secret and typically for relatively low wages. Until quite recently, it was rare for platforms to acknowledge they employed moderators at all. But pressure from journalists, leaks, scandals, and court cases made it impossible for them to deny this work anymore. Now moderators' work is sometimes even mentioned as being mission-critical and essential—words which took on new meaning during the pandemic.

Still, most of these employees are subject to strict nondisclosure agreements (NDAs), and it is mainly because some workers have violated those NDAs that we have the information we do about their working conditions. Sarah Roberts (2019) documents these working conditions through interviews, and one aspect worth highlighting is how these employees work alongside the automated moderation systems discussed above. Roberts (2019, 160) writes:

> Despite word filters that automatically removed specific offensive words on the platforms and sites she moderated, users constantly found ways around the system, using inventive methods to post banned words, or using creative permutations of racial epithets that Melinda [a pseudonym] had to even periodically look up to recognize."

Throughout, Roberts stresses "the need for commercial content moderation to be a human process, or to at least have significant human oversight" (160). That is, even automated systems with relatively simple goals require human supervision, and that supervision is often a morally charged role.

Perhaps more disturbing, though, is the fact that exposure to CSAM, videos and images of violence, hate speech of extreme viciousness, and much more is all in a day's work for these moderators. In his "Trauma Floor" series, journalist Casey Newton (2019) describes one new employee who is tasked with moderating a Facebook post in front of her fellow trainees. "When it's her turn," he writes,

---

Perrigo (2022), among others. See also documentaries from Block and Riesewieck (2018) and Cassidy and Chen (2017).

[15] According to Cassidy and Chen (2017), there are more individuals hired as moderators via third-party contractors than all the official employees of Google or Facebook. Estimates of how many moderators Facebook employs range from fifteen thousand to forty thousand.

> she walks to the front of the room, where a monitor displays a video that has been posted to the world's largest social network. None of the trainees have seen it before, Chloe [a [pseudonym] included. She presses play.
>
> The video depicts a man being murdered. Someone is stabbing him, dozens of times, while he screams and begs for his life.

This, obviously, is not easy work. But while it is clearly psychologically challenging and mission-critical, it is also often undervalued and disrespected. Content moderators, if they are employed directly by the platform company itself, tend to make a fraction of what most other employees on the engineering or business side of things make, with much less job security. More often, though, they are hired via contractors, which the major platforms use to shield themselves from a variety of liability issues. This occurs even when they work on-site with the "regular" employees, who have access to perks and benefits moderators do not. But many moderators work at facilities outside the US—this follows partly from the need for moderators with knowledge of local languages.[16]

No matter their geographic location, though, I want to highlight how these moderators are often harmed by their exposure to an avalanche of toxic content. To me, this is uncontroversial. Reflecting on the type of content they view each day shows that their work opens them up to psychological harms, with long-lasting consequences in some cases. Indeed, one reason we're aware of their working conditions is because of some successful lawsuits filed by former moderators that accused their employers of failing to protect them from harms like posttraumatic stress disorder.[17] It is nasty, psychologically dangerous work.[18]

Just as significant as the fact that these harms exist is the fact that Big Tech tries to conceal them. As noted in the previous section, Big Tech firms' preferred strategy is to instead foreground their efforts to solve the problems of content moderation from a technological lens—specifically, automation through AI. The claim

---

[16] See Perrigo (2022) for reporting on what they call "Facebook's African Sweatshop." While the "sweatshop" moniker may seem off base, other researchers agree. For example, Alfano, Sullivan, and Fard (2022) write, "It would not be unreasonable to compare the crowd-sourced work done as part of the enrichment of NLP datasets to sweatshop labor."

[17] See, for example, Newton (2020a) and Garcia (2018).

[18] And is acknowledge as such by the platform companies. As Perrigo (2022) notes, "Many Facebook content moderators employed by the outsourcing firm Accenture are now asked to sign a waiver before they begin their jobs, acknowledging that they may develop PTSD and other mental health disorders."

I now want to make is that this serves a propagandistic function that, given the severity of the problems at issue, should be uncovered.

3.1. The Propaganda of Automation

While many people express hopes or fears that robots will soon eliminate nearly every job, the actual relationship between technology and human work is summed up in a report published by the research institute Data & Society. "Automated and AI technologies," the authors write, "tend to mask the human labor that allows them to be fully integrated into a social context while profoundly changing the conditions and quality of labor that is at stake" (Mateescu and Elish 2019, 4). Because these technologies mask rather than replace labour, critics have used the derisive names "ghost work" (Gray and Suri 2019), "Potemkin AI" (Sadowski 2018), and "fauxtomation" (Taylor 2018) to describe the current situation of many AI and automated systems in the workplace and the human labor they help conceal.

For example, Astra Taylor (2018) argues that this fauxtomation "reinforces the perception that work has no value if it is unpaid and acclimates us to the idea that one day we won't be needed." All the apps, APIs (application programming interfaces), and even "self-driving drones" that rely on invisible human labour increase the impression that services can simply be rendered, decisions can efficiently be made, and work can just be done, without anyone actually working—despite the fact that this perception is based on an illusion. This, Taylor argues, exists in a long history of capital devaluing labour by implying it isn't needed. "Capitalism lives and grows," she writes, invoking the socialist feminist tradition, "by concealing certain kinds of work, refusing to pay for it, and pretending it's not, in fact, work at all" (Taylor 2018).

We see this dynamic at play in Big Tech's discussion around content moderation—even when we acknowledge that moderation does in fact involve sophisticated AI. Moderators are hidden by outsourcing, third-party contracting, and NDAs. And while platforms rely on the services these moderators provide to make their sites "safe for work," they pay as little as they can get away with.[19] This follows a general pattern within Big Tech, where entire industries rely on services like Amazon's "Mechanical Turk" (MTurk), which the company playfully calls "artificial artificial intelligence." More critically, researchers Mary Gray and Sidd Suri (2019) label this "ghost work." Under these working conditions, workers vie for the chance to complete discreet "micro-jobs," where pay is more often calculated in cents than dollars.[20] A platform like MTurk plays two roles at once. It provides companies with

[19] See Perrigo (2022) for reporting on the pay African moderators receive, along with alleged suppression of the right to unionize by the contractor Meta employs.
[20] See Jones (2021b) for another exploration of the global micro-workforce.

access to a massive pool of cheap labour and does so in a way that keeps these workers at a distance, concealed behind the machine. "The platform and its interfaces," Sadowski (2018) writes, "allow employers to command people as though they were simply operating a mindless machine." A recent report discussed how tech firms often actively seek out displaced populations in refugee camps to perform labelling tasks for pennies (Jones 2021a). To be clear, content moderation is usually done in-house or by contractors—not using MTurk—but the AI systems used to support this work, and that are touted as able to replace this work, are implicated in the larger world of click-work.

According to Taylor (2018), because we often don't know and don't care to know how "automated" services work, "we often believe the hype, giving automation more credit than it's actually due. In the process, we fail to see—and to value—the labor of our fellow human beings." Social media companies like Facebook are happy to endorse this naivety, as well. The notion that AI can satisfactorily address the many problems that currently plague platforms is, in my view, both incredulously optimistic and, I will now add, obfuscating. At best, it tells only a partial story, and one that serves the interest of the technology companies. As Gillespie (2020, 1) says,

> The claim that moderation at scale requires AI is a discursive justification for putting certain specific articulations into place—like hiring more human moderators, so as to produce training data, so as to later replace those moderators with AI. In the same breath, other approaches are dispensed with, as are any deeper interrogations of the capitalist, "growth at all costs" imperative that fuels these massive platforms in the first place.

It is in this way that we should understand Big Tech's public statements about AI as a form of propaganda that serves to redirect concerns from the harms their platforms are implicated in to "solutions" that are not now up to the task and may never be. Moreover, these claims do double duty, as they not only downplay the culpable actions the platforms take—positioning AI as the solution to, rather than source of, the problem—but also erase the work of existing moderators, as well as the harms they experience. In this way, these claims function as a type of *dehumanizing* propaganda, in that they shift attention from the work (and harms) of human beings to that of automated systems. One helpful model of propaganda that illuminates this function is Jason Stanley's conception of "undermining propaganda," which he defines as "a contribution to public discourse that is presented as an embodiment of certain ideals, yet is of a kind that tends to erode those very ideals" (Stanley 2015, 53). In other words, Stanley is interested in cases where specific contributions s*uperficially* appear to be reasonable for a certain discourse, but in fact

undermine that discourse by cutting off, obfuscating, or silencing further debate. In this case, AI is invoked as a solution to pressing moral challenges, but this AI-solutionism in fact undermines our approach to these challenges as it turns a moral question into a technological one. This is *the function* of this rhetoric, and following Stanley's conception, this is irrespective of the intentions with which these suggestions are offered. Mark Zuckerberg may honestly believe in the power of AI, but that does not determine the impact his assertions have on this topic.

In addition to concealing the difficult, often harmful work that moderators do, the emphasis on AI plays another role worth considering. By presenting the decisions as those of a computer system—one powered by AI, no less—algorithmic moderation offers these platforms a shield of neutrality, even objectivity. Beyond being a "black box" that cannot be probed, the turn to AI allows platforms to present these systems as endowed "with powers of objectivity, neutrality, authority, efficiency, and other desirable attributes and outcomes" (Sadowski 2018). This is incredibly valuable to platform companies who aspire to constant growth. And rather than deny the existence of "algorithmic bias," in which users with certain identifiable characteristics are unfairly treated differently by the automated systems than others, Big Tech has instead shown their willingness to adopt an almost single-minded focus on addressing this one particular algorithmic harm (Hao 2021b). This is, in my view, mostly because addressing this problem does not itself challenge the notion that we are concerned with technological problems calling for technological solutions. Moreover, the solutions to these issues often require Big Tech to collect more rather than less data and to exert more rather than less power.

Both tendencies deserve scrutiny. Technology is not itself neutral nor objective, and Big Tech companies should not dictate the scope of the problems of algorithmic justice themselves. And of course, the harms experienced by the moderators employed by platform companies are excluded from this narrow framework of "algorithmic harm." While this harm is not one that is vindictively intended by anyone, it is nonetheless a predictable result. It is part of the background conditions that shape social media, and it is structured by the business decisions of giant, for-profit entities. Those decisions play some role in emphasizing automated moderation while downplaying moderators.

Under capitalism, "automation has an ideological function as well as a technological dimension" (Taylor 2018). The psychological harms experienced by content moderators are not compatible with this ideology, however, and so are erased. It is no surprise that platform companies will use the tools of NDAs, limited third-party liability, and the rhetorical shine of AI to conceal their relationship to harms that befall others. But justice in this domain demands they take responsibility. I argued that we can reasonably describe this rhetoric as propaganda because it functions both to dehumanize existing workers and also to obfuscate and undermine

the debate over responsibility that is urgently needed. I will now address, in my conclusion, how this debate ought to be pursed.

Conclusion

In this paper I offered a partial defense of the claim that AI, especially as deployed by large platforms like Facebook and YouTube, plays some role in increasing violent extremism. I also examined the response these charges have generated from the platform companies, specifically the notion that AI itself might solve these issues. The fact that AI appears as both the problem and the solution may appear odd, but as Gillespie (2020, 2) puts it, "This link between platforms, moderation, and AI is quickly becoming self-fulfilling: platforms have reached a scale where only AI solutions seem viable; AI solutions allow platforms to grow further." As such, the platform companies have a strong interest in vindicating their (future) use of AI even while admitting to its (current) harmful effects. My contribution here has been to demonstrate that in addition to the technical challenges that would need to be overcome for AI to fulfil these dreams, there are also other pressures at play—notably, the propagandistic rhetoric of Big Tech that makes addressing these issues more challenging than they already are. That is, Big Tech has a related interest in marginalizing and erasing much of the human labour that supports their AI, especially regarding content moderation, as it does not fit their narrative of continuous technological progress.

One important aspect of this debate that I have not addressed in this paper, but that requires elaboration in future work, is the fact that algorithmic moderation may serve as a solution to imposing these harms on content moderators. That is, "the strongest argument for the automation of content moderation may be that, given the human costs, there is simply no other ethical way to do it, even if it is done poorly" (Gillespie 2020, 4). This may be true, and whether it is feasible is largely an empirical question. However, it is a question that is itself obscured by Big Tech's rhetoric on this topic.

Furthermore, given that total automation is likely impossible, we still must grapple with the hard question of whose interests are ultimately served by these systems. As Newton (2020b) puts it, there is an urgent trade-off in content moderation: "the use of automated systems that are error-prone but invincible, versus the use of human beings who are much more skilled but vulnerable to the effects of the job." Automated moderation generally is overzealous, removes too much "good speech," and frustrates users. But human moderation has the clear cost of debilitating mental health conditions. "So far," Newton writes, "no global-scale technology company has managed to get this balance right. In fact, we *still have no real agreement on what getting it 'right' would even look like*" (2020b; emphasis added). This may be true, but my point here is that we can only approach this topic

adequately when we acknowledge there is a trade-off to be made, which is incompatible with the assertation that better AI can handle everything itself.

The massive scale of social media presents another problem, with which I will close. That problem is the distribution of responsibility in what is a global system with billions of moral agents is complicated by the fact that the users of social media are integral to the platform companies' products. As social media users, we all likely share part of the responsibility for the harms of platform companies generate, at least in so far as our actions provide support for their practices. We are therefore complicit in the injustices discussed above. However, as Iris Marion Young (2006) argues, shared responsibility is not always shared evenly.[21] She articulates four "parameters of reasoning" that allow us to make headway on the issue of distributing responsibility when faced with structural injustice within her social connection model of responsibility.[22] These are *power*, which is influence over the processes that produce these outcomes; *privilege*, the ability to adapt without suffering serious deprivation; *interest*, both in maintaining and in transforming given structures; and *collective ability*, the organizational capacity to address a particular issue (Young 2006, 127–30). Each of these parameters could inform our approach to addressing the harms outlined above, and each deserves more space than I can allot here. But I will highlight two features of this model that demonstrate the insights available. The first concerns power, and the fact that "the agents with the greatest power within social structures often have a vested interest in maintaining them as they are," which Young takes to imply that "external pressure on the powerful is often necessary to move these agents to action, and to prevent them from taking superficial steps rather than making serious changes" (127). It is clear that this is the case with platform companies—who are some of the richest organizations in the world—and the harms they produce. This is compounded by the fact that the subjects of the harms I've discussed here—content moderators, victims of extremist violence, and even some (potentially) radicalized individuals themselves—are more marginalized and less powerful than those in Silicon Valley. While responsibility falls on many different agents in a distributed system, we must not ignore who predominantly holds the levers of power. And putting pressure on these powerful agents is undoubtedly one avenue through which regular users can and should discharge their responsibility.

---

[21] As she says: "Different agents plausibly have different kinds of responsibilities in relation to particular issues of justice, and some arguably have a greater degree of responsibility than others" (Young 2006, 126).

[22] This social connection model of responsibility, she says, is meant to "correspond to the intuition that those who participate by their actions in the structural processes that produce injustice bear some responsibility for correcting this injustice," (Young 2006, 125) and thus seems an apt model for the current discussion.

This takes me to the second insight from Young I wish to highlight, which is that while powerful and privileged actors are likely *more* responsible for change, she maintains that "victims of injustice share responsibility with others for cooperating in projects to undermine the injustice" (128). This is because, quite simply, "victims of injustice have the greatest interest in its elimination, and often have unique insights into its social sources and the probable effects of proposals for change" (128). In the case of platforms and content moderation, this implies not only that moderators themselves have some responsibility to improve their conditions but also that they can and should shape the direction that those improvements take.[23] Concretely, this should mean that platform companies (and the companies they contract with) ought not to suppress workers from organizing, should seek moderators' feedback on policy decision, and should offer potential paths for advancement within the organization. They should, moreover, uncover the moderators' opinions on how to implement automation in their workplace.

The problems discussed above are complex, and their solutions are not obvious. But given that social media platforms are at the centre of a great number of increasingly significant moral problems, it is imperative that we approach this task with urgency, care, and a full appreciation of the harms at stake. This requires seeing beyond the propagandistic rhetoric of Big Tech and instead asking how current harms can be properly acknowledged, mitigated, and compensated.

References

Alfano, Mark, J. Adam Carter, and Marc Cheong. 2018. "Technological Seduction and Self-Radicalization." *Journal of the American Philosophical Association* 4, no. 3 (Fall): 298–322. https://doi.org/10.1017/apa.2018.27.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2021. "Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System." *Synthese* 199, no. 1–2 (December): 835–58. https://doi.org/10.1007/s11229-020-02724-x.

Alfano, Mark, Emily Sullivan, and Amir Ebrahimi Fard. 2022. "Ethical Pitfalls for Natural Language Processing in Psychology." In *Handbook of Language Analysis in Psychology*, edited by Morteza Dehghani and Ryan L. Boyd, 511–30. New York: Guilford Press.

[23] It is important to note that, for Young, "victims of injustice have a responsibility to work together to improve their situation, but they are unlikely to succeed without the help and support of other less-vulnerable people who make industry behavior public and who pressure companies to change policies or restructure their business relationships" (129).

Ball-Burack, Ari, Michelle Seng Ah Lee, Jennifer Cobbe, Jatinder Singh. 2021. "Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection." In *FAccT '21: Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, 116–28. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445875.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Block, Hans, and Moritz Riesewieck, directors. 2018. *The Cleaners*. Gebrueder Beetz Filmproduktion. 88 min.

Cassidy, Ciarán, and Adrian Chen, directors. 2017. *The Moderators*. Field of Vision. 20 min. https://fieldofvision.org/the-moderators.

DiResta, Renée. 2020. "The Right's Disinformation Machine Is Getting Ready for Trump to Lose." *Atlantic*, October 20, 2020. https://www.theatlantic.com/ideas/archive/2020/10/the-rights-disinformation-machine-is-hedging-its-bets/616761/.

Dwoskin, Elizabeth, and Craig Timberg. 2021. "Misinformation Dropped Dramatically the Week after Twitter Banned Trump and Some Allies." *Washington Post*, January 16, 2021. https://www.washingtonpost.com/technology/2021/01/16/misinformation-trump-twitter/.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Garcia, Sandra E. 2018. "Ex-Content Moderator Sues Facebook, Saying Violent Images Caused Her PTSD." *New York Times*, September 25, 2018. https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html.

Geiger, R. Stuart. 2016. "Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space." *Information, Communication & Society* 19 (6): 787–803. https://doi.org/10.1080/1369118X.2016.1153700.

Gessen, Masha. 2018. "Why the Tree of Life Shooter Was Fixated on the Hebrew Immigrant Aid Society." *New Yorker*, Oct 27, 2018. https://www.newyorker.com/news/our-columnists/why-the-tree-of-life-shooter-was-fixated-on-the-hebrew-immigrant-aid-society.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

———. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7, no. 2 (July). https://doi.org/10.1177/2053951720943234.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7, no. 1 (January–June). https://doi.org/10.1177/2053951719897945.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Hao, Karen. 2021a. "How Facebook Got Addicted to Spreading Misinformation." *MIT Technology Review*, March 11, 2021, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

———. 2021b. "The Race to Understand the Exhilarating, Dangerous World of Language AI." *MIT Technology Review*, May 20, 2021. https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/.

Jones, Phil. 2021a. "Refugees Help Power Machine Learning Advances at Microsoft, Facebook, and Amazon." *Rest of World*, September 22, 2021. https://restofworld.org/2021/refugees-machine-learning-big-tech/.

———. 2021b. *Work without the Worker: Labour in the Age of Platform Capitalism*. London: Verso.

Lewis, Becca. 2020. "I Warned in 2018 YouTube Was Fueling Far-Right Extremism. Here's What the Platform Should Be Doing." *Guardian*, December 11, 2020. https://amp.theguardian.com/technology/2020/dec/11/youtube-islamophobia-christchurch-shooter-hate-speech.

Lewis, Rebecca. 2018. *Alternative Influence: Broadcasting the Reactionary Right on YouTube.* Data & Society Research Institute. https://datasociety.net/library/alternative-influence/.

Lynch, Michael Patrick. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York: Liveright Publishing.

Mateescu, Alexandra, and Madeleine Clare Elish. 2019. *AI in Context: The Labor of Integrating New Technologies*. Data & Society Research Institute. https://datasociety.net/library/ai-in-context/.

McCrosky, Jesse, and Brandi Geurkink. 2021. *YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm*. Mozilla Foundation. https://mzl.la/regrets-research.

Metz, Cade, and Mike Isaac. 2019. "Facebook's A.I. Whiz Now Faces the Task of Cleaning It Up. Sometimes That Brings Him to Tears." *New York Times*, May 17, 2019. https://www.nytimes.com/2019/05/17/technology/facebook-ai-schroepfer.html.

Michaelson, Eliot, Jessica Pepp, and Rachel Sterken. 2021. "Online Communication." *Philosopher's Magazine*, August 17, 2021. https://www.philosophersmag.com/essays/249-online-communication.

Newton, Casey. 2019. “The Trauma Floor: the Secret Lives of Facebook Moderators in America.” *Verge*, February 25, 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

———. 2020a. “Facebook Will Pay $52 Million in Settlement with Moderators Who Developed PTSD on the Job.” *Verge*, May 12, 2020. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

———. 2020b. “YouTube Gets Sued by Its Moderators.” *Interface*, no. 572, September 22, 2020. https://www.getrevue.co/profile/caseynewton/issues/youtube-gets-sued-by-its-moderators-280023.

Noble, Safya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Perrigo, Billy. 2022. “Inside Facebook’s African Sweatshop.” *Time*, 14 February, 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Sadowski, Jathan. 2018. “Potemkin AI.” *Real Life*, August 6, 2018. https://reallifemag.com/potemkin-ai/.

Smith, Laura. 2021. “Lone Wolves Connected Online: A History of Modern White Supremacy.” *New York Times*, January 26, 2021. https://www.nytimes.com/2021/01/26/us/louis-beam-white-supremacy-internet.html.

Solon, Olivia. 2018. “The Rise of ‘Pseudo-AI’: How Tech Firms Quietly Use Humans to Do Bots’ Work. *The Guardian*, July 6, 2018. https://www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies.

Solsman, Joan E. 2018. “YouTube's AI Is the Puppet Master over Most of What You Watch.” *CNET*, January 10, 2018, https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.

Sonderby, Chris. 2019. “Update on New Zealand.” *Facebook Newsroom*, March 18, 2019. https://perma.cc/ZA85-2Y3X.

Stanley, Jason. 2015. *How Propaganda Works*. Princeton, NJ: Princeton University Press.

Statt, Nick. 2014. “Zuckerberg: ‘Move Fast and Break Things’ Isn't How Facebook Operates Anymore.” CNET, April 30, 2014. https://www.cnet.com/tech/mobile/zuckerberg-move-fast-and-break-things-isnt-how-we-operate-anymore/.

Stecklow, Steve. 2018. "Hatebook: Inside Facebook's Myanmar Operation; Why Facebook Is Losing the War on Hate Speech in Myanmar." *Reuters*, August 15. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate#article-hatebook.

Taylor, Astra. 2018. "The Automation Charade." *Logic,* no. 5, August 1, 2018. https://logicmag.io/failure/the-automation-charade/.

Tubaro, Paola, Antonio A. Casilli, and Marion Coville. "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence." *Big Data & Society* 7, no. 1 (January–June). https://doi.org/10.1177/2053951720919776.

Warzel, Charlie. 2019. "Mass Shootings Have Become a Sickening Meme." *New York Times*, April 28, 2019. https://www.nytimes.com/2019/04/28/opinion/poway-synagogue-shooting-meme.html.

York, Jillian C. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. New York: Verso.

Young, Iris Marion. 2006. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy & Policy* 23, no. 1 (January): 102–30.

Zadrozny, Brandy. 2021. "YouTube's Recommendations Still Push Harmful Videos, Crowdsourced Study Finds." *NBC News*, July 7, 2021. https://www.nbcnews.com/news/amp/rcna1355.

Zimmermann, Annette, Elena Di Rosa, and Hochan Kim. 2020. "Technology Can't Fix Algorithmic Injustice." *Boston Review*, January 9, 2020. https://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.

MICHAEL RANDALL BARNES is a postdoctoral associate with the Humanising Machine Intelligence project at the Australian National University. He works mainly on issues relating to how speech harms, and is particularly interested in the harms made possible by online speech.