Accelerated Epistemic Harm: Understanding the Role of Social Media Engagement Algorithms in Online Radicalization

by

Liam Burke

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

Carleton University
Ottawa, Ontario

Abstract

Social media use appears to play a role in radicalizing an increasing number of people. I problematize what I characterize as one popular picture of radicalization, which demands that an agent be socially isolated, have mental health issues, a propensity to violence, and becomes radicalized on a timeline. I argue that what is particularly concerning about online radicalization, versus offline radicalization, is the accelerated epistemic harm that prolonged social media use and exposure to its engagement algorithms can cause to the epistemic capacities of the agent, drawing from Fricker's concept of epistemic injustice. I then argue that social media companies share some responsibility for online radicalization, having fostered an environment that can cause epistemic harm. I conclude by sketching ways we might combat this environment at the level of the agent and community, problematizing the virtue ethics-based approach that is common in such recommendations and instead favouring a legislative approach.

Acknowledgements

I am grateful to the many excellent professors at Carleton University with whom I have studied throughout my time as a philosophy student. Significant input and guidance for this thesis was provided by Gabriele Contessa, Gordon Davis, Christine Koggel and David Matheson. Special thanks is due in particular to my supervisor, David Matheson, without whose care and tireless effort this thesis would doubtless have not come to life.

Special thanks are also due to my family, for my life, and to my partner, Manahil, for care, love, and art, and for helping me drag this project kicking and screaming into the light.

Table of Contents

“Everything we hear is an opinion, not a fact. Everything we see is a perspective, not the truth.”

-Marcus Aurelius (misattributed)

“My mind is made up – don't confuse me with facts.”

-United States Congressman Earl Landgrebe (also misattributed)

Introduction

The shifting needle

On the evening of January 29, 2017, 27-year-old Alexandre Bissonnette entered the Islamic Cultural Centre of Quebec City and opened fire with a pistol, killing six and seriously wounding five. Friends of Bissonnette said he had far-right, white nationalist, misogynistic and Islamophobic views. Often, Bissonnette would take to Facebook to engage in online trolling – an activity where Internet users make inflammatory statements to garner a reaction – where he would post hateful diatribes against immigrants and feminists. He was an outspoken supporter of Donald Trump and Marine Le Pen and a voracious consumer of Ben Shapiro's Twitter feed.

We have been conditioned to think of people radicalized online as being like Bissonnette – social misfits who grew to become antisocial pariahs, dwelling in increasingly hateful online meeting places before lashing out in one terrifying act of violent extremism. However, evidence is shifting the needle away from this popular picture of radicalization to broaden the types of people we should consider as radicalized.

As I write this, I sit in my basement apartment in Kanata, quite a distance away from the downtown core of Ottawa, which has been unlawfully occupied for weeks by a convoy of disenfranchised truckers and blue-collar workers. The convoy was organized by Pat King, an outspoken white supremacist. Confederate and Nazi flags were spotted at convoy rallies in Ottawa – yet the convoy was not overtly a white supremacist rally, and not every member of the convoy appeared to be an outspoken white supremacist like Pat King. Of course it would be, at best, facile and insulting to assume or assert that every

trucker and blue-collar worker present is an overt white supremacist. So why did so many rally behind him?

The sentiment for the trucker convoy grew behind a slow burn of two years of COVID denial, anti-vaccine rhetoric, pandemic disinformation, and refusal to wear a mask and to abide by distancing guidelines. Somehow, thousands of people managed to convince themselves that all of these reasonable civic duties imposed too strongly on their fundamental Canadian freedoms, enough to drive across the country behind a vocal racist to harass and terrorize the capital city of their country. Many participated in psychological warfare, blaring truck and train horns around the clock for weeks on end. Other convoy members intimidated mask wearers, food banks, and the homeless; several attempted arson and mass murder against the residents of an apartment building.

A 2012 report from the United States-based think tank the Bipartisan Policy Center's Homeland Security Project states, "We believe that this trend [of the Internet radicalizing people] will continue and that future terrorist attacks against the United States and its interests will involve individuals who have been radicalized – at least in part – on the Internet." (1)

A popular picture of radicalization

As this thesis will explore the effects of Internet radicalization, it behooves us to first explore a working characterization of radicalization before examining how it relates to the Internet. As I will understand it, radicalization is a discursive process that changes the viewpoints of an agent to one that is extreme in a number of ways.

First, radicalization predisposes agents to violence, either physical or ideological. Physical violence is a broad spectrum ranging from mass shootings, arson and physical assault to intimidation, such as marches, harassment, displaying of hate symbols, or cross burning. Ideological violence is found when an agent harbours hateful views, either secretly or publicly, such as racism, anti-Semitism, or Islamophobia. It is important to note, however, that a radicalized agent can become predisposed to ideological violence without ever participating in physical violence, and I believe it is plausible that it is a great deal more common for agents to become predisposed to ideological violence than to physical.

Second, radicalized agents become distrustful of canonical ways of gathering belief. Radical views are ones that reject traditional sources of testimonial knowledge, such as science and policy experts in government, or even reputable bias-free websites such as Snopes.com. Radicalization forces agents to reject sources of evidence that render their views demonstrably false by following increasingly bizarre and far-fetched conspiracy theories: by believing, for instance, that Snopes.com is owned by billionaire George Soros, or that Democrat politicians such as Hillary Clinton are part of a pedophile cult.

As they burn their bridges, so to speak, with what we consider to be reputable ways of gathering belief, radicalized agents place themselves in an increasingly precarious epistemic situation, such that when acquiring new knowledge, they look only to the testimonial sources that have radicalized them, or similar agreeable ones in the same testimonial “family”.

The Canada Centre for Community Engagement and Prevention of Violence posits radicalization as a system that happens by degrees, where one first eschews mainstream beliefs and may gradually move towards violence. It defines three separate components of this process: radicalization, radicalization to violence, and finally violent extremism:

> "Radicalization is a process by which an individual or a group gradually adopts extreme positions or ideologies that are opposed to the status quo and challenge mainstream ideas.
>
> Radicalization to violence is the process by which individuals and groups adopt an ideology and/or belief system that justifies the use of violence in order to advance their cause.
>
> Violent extremism is a term describing the beliefs and actions of people who support or use violence to achieve extreme ideological, religious or political goals." (7)

The document implies a "timeline" where a process of radicalization may be teleological, in that it may eventually push an agent towards violent extremism. The same document dissects the various behaviours agents who have become radicalized engage in, which also covers behaviours they may engage in if they are not committed to violent extremism:

> "Radicalization to violence occurs when a person or group takes on extreme ideas and begins to think they should use violence to support or advance their ideas or beliefs. These beliefs can fall along a wide spectrum of ideologies, including political and religious ideologies[...] While violent attacks are the most extreme result of radicalization to violence, Canadians engage in a range of non-violent behaviours, both offline and online, in support of violent extremism:
>
> • Funding violent extremist or terrorist groups.
>
> • Spreading messages and narratives that incite violence and hatred.
>
> • Recruiting individuals to become part of a violent extremist or terrorist group.
>
> • Travelling to join violent extremists or terrorist groups abroad.
>
> • Expressing support for a terrorist group." (12)

It is clear that our popular picture of radicalization, exemplified by such documents as the above, conceives of radicalized agents of having a propensity towards at least ideological if not physical violence. The process of radicalization requires vulnerable agents; in other words, angry people who need a sense of belonging. It also needs an agenda – a manifesto of some kind to make radicalized agents feel justified in directing violence at others. Consonantly with the popular conception of radicalization I describe, angry and isolated young men without direction in their lives are the most vulnerable to radicalization. But, increasingly, and I will argue due in no small part to Internet algorithms, more people outside that precarious bubble are falling prey to misinformation and becoming radicalized.

The specifics of online radicalization

Radicalization is not a new phenomenon – it is in fact much older than the Internet. An article on online radicalization by the World Economic Forum points out that Gavrilo Princip, infamous assassin of the archduke Franz Ferdinand, was radicalized by state media (Letzing and Berkley). In a similar vein, Timothy McVeigh purchased and wore merchandise from the Ku Klux Klan while in the army, grew obsessed with the Branch Davidian compound in Waco, Texas, and became increasingly anti-government before bombing the Alfred P. Murrah federal building in Oklahoma City in 1995, long before the World Wide Web had become the vast, established entity it is today. So what exactly is so special about online radicalization?

Our first impression might be to think that online radicalization worsens the epistemic poverty faced by a radicalized person simply because it provides easy access to discreditable epistemic sources. Instead of being limited to hateful state propaganda, or fringe print media, there are all sorts of places to turn online now and be faced with content ranging from questionable, borderline jokes and comments to outright hate speech.

Neumann (2013) borrows from Weimann in distinguishing between radicals' *instrumental* and *communicative* uses of the Internet. Instrumental uses are related to common, everyday Internet usage: radicals use search engines, email and book travel or look up pictures or maps. They may also use the Internet to raise money for their causes via crowdsourcing websites. In this manner, asserts Neumann, radicalized people and violent extremists don't use the Internet that much differently from everybody else.

Communicative uses, on the other hand, grow increasingly relevant and prevalent in today's online architecture, as the way we keep in touch with, debate with, and learn from one another changes constantly. Violent extremists – and everyone else, for that matter – have shifted away from using static websites and bulletin boards to communicating on social media.

Part of the nature of content on social media is that creators can code their messages in memes, iterative in-jokes that evolve and change meaning over time. Videos on TikTok and YouTube with questionable, borderline hateful content don't need to pull back the curtain right away. They can just show a meme associated with racist or misogynist content, as a punchline to a joke, and the viewer's mind can fill in the blank. This also has the effect of not making the viewer have to read or repeat an outright objectionable statement, something that might make them think twice about what they read.

The biggest and most important difference from the static websites and bulletin boards of the 90s and early 2000s, however, is that agents no longer need to go actively looking for hateful or objectionable content. In earlier days, agents who became radicalized had to first express interest in extremism, and then had to learn a website address from another source – like knowing the password for a speakeasy. Now agents can fall into misinformation rabbit holes and be recruited into extremist movements on social media, all without ever seeking it out.

Online engagement algorithms have a compounding effect. Simply put, when you engage with content on a social media website – by liking, commenting, or just by viewing it – you help build a data profile of yourself for the website, ensuring you will

see similar content in the future. This compounding effect accelerates radicalization in unpredictable ways. Our use patterns on social media – what we watch, like and comment on – yields us more of the same, and contrary opinions slowly shrink in our proverbial rear view mirrors.

Shifting the focus

As we can see, our first impression is lacking. The problem of online radicalization runs deeper than just providing vulnerable agents easy access to various discreditable epistemic sources. Consonantly both with the popular picture of radicalization I have sketched and the Bipartisan Homeland Center's report, during a process of radicalization, agents can often become distrustful of creditable epistemic sources, often embracing discreditable ones instead. Thus, a process of radicalization does not merely increase the *supply* of discreditable sources to an agent – it increases the demand as well. Agents who become radicalized also tend to base their beliefs on things that "feel" correct, or that have the quality of "truthiness", as coined by Stephen Colbert. Truthiness in a belief appeals to an agent's perception or intuitions, rather than logical soundness, rationality or evidence (Alfano 2009).

Throughout this thesis, it will behoove us to keep in mind that social media algorithms are created from the ground up with the ultimate goal to maximize profit for the companies that create them. Part of how they reach that goal is ensuring you are put in touch with the most popular content on the platform. To this end, content that is controversial, because it says forceful or objectionable things, is more likely to show up in your feed because of the disproportionately high amount of engagement it receives

(McCluskey 2021). This is how controversial influencers such as Joe Rogan and Jordan Peterson show up in most everyone's recommended feed eventually – if you don't engage with it, someone else whose content you like has. And actually clicking through on it, signifying to the algorithm that you have engaged with this content, often results in a laser focus on more controversial content, and a never-ending cascade of similarly-minded content.

Thus, we see that online radicalization has changed significantly in the past ten years as social media and its related algorithms have gained traction. Extremist messages have an unimaginably wider reach and audience; orders of magnitude larger than where they were just ten years ago.

By discussing the wider reach and audience that extremist messages enjoy in today's world, I hope to show that our way of thinking of radicalization is outmoded. We typically think of radicalized violence as involving one unbalanced agent, typically battling some degree of chronic mental illness, who actively seeks out and engages with increasingly hateful content until they are finally pushed to acts of terrorism. This is because we think of hateful views as irrational, and the only agents who collect and harbour them to be irrational ones.

Instead, I characterize the current picture online as one where racists, misogynists, and other people who harbour hateful views are all comparing notes and rubbing shoulders together. Alfano, Carter and Cheong, in "Technological Seduction and Self-Radicalization" (2018), write: "Surrounding the dark shadow of lone-wolf terrorism is the broader penumbra of self-radicalization that results in less dramatic but still worrisome actions and attitudes. The 'Unite the Right' rally in Charlottesville, Virginia in

2017 brought together neo-Nazis, white supremacists, and white nationalistic sympathizers for one of the largest in-person hate-themed meetings in the United States in decades. Many of the participants in this rally were organized and recruited via the Internet. Even more broadly, white supremacists and their sympathizers met and organized on r/The_Donald (a Reddit community) and elsewhere during the 2016 American presidential campaign that resulted in the election of Donald Trump." (3)

As radicalized agents grow increasingly vulnerable to misinformation, it becomes more possible to organize the situation we saw unfold in Ottawa: white supremacists can organize trucker convoys not under neo-Nazi or white supremacist banners, but under the guise of a "Freedom Convoy", gathering thousands of people who may not fit the profile of chronically mentally ill social outcasts. And, all the while, these violent extremists are acting as epistemic sources for the members of the convoy: feeding them a steady diet of misinformation and half-truths, riling them up and inciting them to harassment, violence and domestic terrorism. This is the new face of radicalism.

This thesis makes several contributions to the philosophical literature of online radicalization. First, I aim to show that online radicalization, as accelerated by social media algorithms, problematizes a popular picture of radicalization. This popular picture must be amended, I will argue, because it demands of radicalized agents things that may not always be true; for example, that they be mentally ill social outcasts, prone to violence, and that they are radicalized on a discernible timeline.

Second, I aim to adapt Fricker's (2007) conception of epistemic injustice, and epistemic harm, to describe how exactly I believe an agent caught within a process of online radicalization can be harmed in a way that is difficult to detect, report, or predict.

The third contribution is just to show that engagement algorithm-accelerated online radicalization is plausibly distinct from offline radicalization because of the accelerated epistemic harm it can cause to an agent.

Chapter 1: Who Is Involved?

Identifying and collapsing concerned parties

At first blush, it seems we can identify at least three main concerned parties (hereafter referred to as stakeholders) in the problem of online radicalization.

The first concerned party in our equation is the companies that create and monitor the algorithms as well as manage the data used in their operation. Companies of note in this category include:

Alphabet: the parent company of Google and YouTube;

Meta: the parent company of Facebook and Instagram; and

ByteDance Ltd: the owning company of TikTok (known as Douyin in China).

The role of content creation does not fall to the company managing the platform and algorithm, however. Social media platforms are somewhat unique inventions of the Internet age, in that the content available for consumption by the audience is, by and large, created by that audience itself.

This model is far from being perfectly symmetrical, however; the lion's share of audience attention on social media goes to the most successful users and paid content by businesses, as a form of advertising. Over the course of a social media application's lifecycle, content creation becomes effectively outsourced to "influencers". Influencers are content creators who have garnered a large audience and a high degree of clout on the social media platform in question. They write, post pictures and make videos for consumption by other social media platform users. Many influencers make their living off of their social media presence, through business sponsorships and monetary compensation awarded per view by the social media company.

At the bottom of the proverbial food chain, we have the average end-user, the intended audience of platform content. She is the person responsible for contributing to a post's total views, or by interacting with it by "liking" it, clicking a button to demonstrate her approval of the content. Again, the user often contributes to content creation herself; more than viewing and liking a post or video, she may comment on it as well, or make one of her own in response. She may, as on Twitter, "retweet" another's post to copy it to her own page and share it with her own audience. She may, as on YouTube, make a playlist of other people's content, making it possible for other users to see at a glance what her interests are, and follow her curated choices accordingly. All of this is to show that, on social media, the distinction between content creator and user is never completely clear.

Sullivan et al., in "Vulnerability in Social Epistemic Networks" (2020) place a stronger emphasis on the plurality of sources from which agents receive information in a modern social environment. This is a point that must be strongly emphasized in social epistemological frameworks in order to properly understand the spread of misinformation and its role in radicalization. The authors put forth the following thought experiment:

Agent A receives all the new information she doesn't already possess about the outside world from agent B. This creates a two-person relationship – a speaker-hearer dyad.

Now suppose there is more to know about the world than A can learn from B. She goes to agents C, D and E in search of new information. All appears to be well, and A seems to flourish in this new diverse epistemic environment.

But now we find out that C, D and E have not acquired all of their beliefs

independently – in fact, they are each reliant on yet another agent for outside knowledge, agent F. So while A appears to get her sources of exterior knowledge from four separate independent networks, she in fact only receives her knowledge from two: agents B and F. Because C, D and E all receive their exterior knowledge from F, they serve to amplify F's perspective threefold, rather than acting as independent sources. This position makes A epistemically vulnerable.

Sullivan et al. argue that in the framework they envision, agents are vulnerable when they appear to be receiving news from a plurality of sources, but in fact each of these sources has received its news from the same, discreditable source. The seeming concurrence of voices can distort an agent's ability to criticize the content they engage with: "Three is better than one, but only when the three have independent evidence, employ different methods of inquiry, or differ in their sensitivity to epistemic reasons." (Sullivan et al. 5)

Case study: TikTok and COVID-19 meme disinformation

Basch et al. (2021), in a study of COVID-19 vaccine-related content on the social media platform TikTok, categorized videos as either discouraging a vaccine or encouraging one. The authors found that there was only a minor difference in their sample, both in number and in numbers of views, likes and comments gathered, between videos that discouraged the vaccine versus videos that encouraged it. However, and more troublingly, of the videos that discouraged the vaccine, a disproportionate amount of the views went to videos that parodied an adverse reaction to the vaccine – a vaccine that was not, at the time of the study, fully developed or yet available to the public. (Basch et

al. 4) The authors believe that this worrying find demonstrated a deliberate effort to promote anti-vaccine sentiment among TikTok users. Basch et al. assert that "[t]he vast majority of young people between the ages of 18 and 30 years of age rely on online sources of health information", and that "[w]hile young people may have high levels of digital literacy, they tend to have comparatively low levels of health literacy, limiting their ability to critically evaluate online content, including that found on social media." (Basch et al. 5)

The authors provide a useful characterization of both *misinformation* and *disinformation*, two terms I believe are commonly used interchangeably but have quite different, if related, meanings. According to Basch et al., "[w]hereas misinformation involves drawing erroneous conclusions about a given phenomenon based on incorrect or incomplete facts, a disinformation campaign involves active effort to spread false information to advance a particular agenda." (Basch et al. 2)

The authors' findings are troubling, as any evidence of a campaign of active disinformation should be. But it should be noted that once the misinformation is out there, it doesn't seem to matter whether an agent intends to mislead or not when they spread it. In other words, a campaign of *disinformation* can quickly become one of *misinformation* on social media; intentionally disingenuous and misleading messages can be reposted, referenced, mentioned or otherwise have their signals boosted by other users, who may have any number of reasons to be less than fastidious about fact-checking that do not constitute malicious intent.

Consider again the role of the influencer. Certainly these people carry more weight in how information flows across the social media platform they use, but it should

be noted that they themselves are users; they too can be vulnerable to compromise of their epistemic structures in the way Sullivan et al. describe. Suppose a number of influencers are vulnerable to a single source, which turns out to be disreputable. These influencers do not need to know they are lying to their followers; they may genuinely believe the message they are boosting is truthful, or they may want it to be true badly enough not to fact check it. In other words, they may not need to actively participate in a campaign of *disinformation* to unwittingly participate in one of *misinformation.* In any case, all of their followers, reliant on them for exterior knowledge just as agent A was, are now vulnerable to this disreputable source of information. This effectively creates a kind of feedback loop, and the longer it goes on, the worse the problem will become. As influencers' popularities rise, so too does the sway their words hold on their platform, as their content becomes more commonly recommended the more engaged with it is. In effect, algorithms develop a "gravity well" around influencers that only deepens over time.

Concerns and potential risks

One might worry whether the picture Sullivan et al. describe also characterizes the way *veridical* information might spread. After all, there is a risk that *all* online engagement will lead to some kind of vulnerable testimonial network. However, I think this concern is easily defused. Networks with limited sources of testimony such as the one Sullivan et al. describe are still *vulnerable* networks, carrying with them a potential risk to the agent. But so long as the information relayed in them is truthful, and verifiable, the risk to the agent remains just that – potential.

When we say *risks* of being in a vulnerable testimonial network, what do we have in mind? An agent could practice "epistemic virtue" and still be in a vulnerable testimonial network. Vulnerable testimonial networks are, by themselves, not sources of harm – we can think of them as "suboptimal architecture", like a sagging house. If these networks happen to transfer truthful and non-trivial information to the agent, so much the better. They do, however, impose certain penalties on the epistemic functions of an agent, and if an agent does not actively work to get out of them, there is some cause for concern. But we have reason to believe, I think, that people who are "epistemically virtuous" will probably get out of them on their own, whereas "epistemically vicious" people are more likely to remain there. Perhaps it's the case, then, that vulnerable testimonial networks are part of a symptom, or act as a force multiplier, of a larger problem. It seems that at least some radicalized agents did not have these sorts of vulnerable testimonial networks, or else radicalization would only be as new as social media. And, conversely, other agents can remain in vulnerable testimonial networks for some time and not become radicalized to violence. The epistemically vicious will still commit epistemic sins such as "wanting to believe" and related cases of flawed thinking.

We should note, again, that vulnerable testimonial networks are a risk to *everyone* who uses social media. We will see in later sections that the nature of social media algorithms is to take note of what content you engage with and filter your view so that, over time, you see more and more of the same type of content. This feature makes "the walls close in", so to speak, and turns every testimonial network into one of "suboptimal architecture". This problem is only exacerbated for those agents who are epistemically vicious.

Perhaps it is enough to say that epistemically virtuous agents will, by their regular courses of action, mitigate the problems of their vulnerable testimonial networks. "Epistemic virtue" is a strategy espoused by several information ethicists, such as Miller & Record (2013), Simpson (2012), and Alfano et al. (2018). Later in this thesis, we will analyze and problematize the use of epistemic virtue and vice as a strategy for individual agents to combat social media radicalization. For now, let us define the epistemically virtuous as those who, by force of habit, question their sources, interrogate their beliefs against new information, and challenge their peers when their peers exhibit flawed thinking or reasoning about the world.

The epistemically vicious, on the other hand, make a weaker effort, or no effort, to mitigate the problems of their vulnerable testimonial networks. They may be unaware or unbothered that their testimonial knowledge comes from a limited number of sources. Or, they may be taken by the illusion, as Sullivan et al. are concerned about, that the number of sources reporting that *p* is sufficient reason to believe in the force of the argument that *p.*

Thus, vulnerable testimonial networks worsen the problem of radicalization by subtly decreasing the amount and variety of sources an agent has access to, while maintaining the appearance of a broad array of sources to amplify a single message. Further, the degree to which individual agents fight to free themselves from a vulnerable testimonial network can also serve as a bellwether to determine whether that particular agent is epistemically virtuous or vicious.

Watching, liking, listening, engaging

In the 20th century, television networks tracked the popularity of their show by using Nielsen ratings, an audience measurement system developed by Nielsen Media Research first used for radio in the 1940s. It is easy to think of the number of views, likes or comments on a social media post, picture or video as a simple set of metrics measuring a video's popularity and engagement, much like the Nielsen rating. Aaron Charles, in an article for Chron's Small Business series, expounds on how liking a video helps its ratings. Charles asserts a given video's number of views and number of likes enjoy a symbiotic relationship, in that more views can lead to more likes and vice versa.

In other words, rather than views, likes and comments being different ways of measuring how popular a video is, these three measurements work alongside one another to provide a detailed analytical picture. Videos with lots of likes or comments on YouTube may be more likely to feature highly in YouTube's recommendation list, thus leading to more views.

In fact, there is even more to liking and commenting than merely showing how engaged with a given social media post is. Comments, for instance, may not be an endorsement of a post or video. Controversial posts may receive a great deal of comments disagreeing with the content in the post. On Twitter, posts with more comments than likes are derisively referred to as having been "ratio"d, or being subject to a disproportionate ratio of comments versus likes. One major limitation of studies such as Basch et al.'s, then, is an inability to measure the discursive element at play – to wit, what exactly is happening in the comments section, and how other users are making their own videos to respond to the ones collected in the study.

Chapter 2: How Do Algorithms Operate?

How algorithms engage users

In much of what follows, the reader will have to allow themselves to be led into the murky bog of speculation. The inner workings of algorithms are kept under proprietary lock-and-key by the companies that own them, since these algorithms are major profit drivers for those companies (see for example Miller and Record 2013). We must thus proceed while projecting against the wall only the broad strokes of what drives companies overall (the pursuit of profit) and specifically those that deal in user experience and communication, such as social media companies (the acquisition of data). Let such projections illuminate the furtive grasping of our guesswork in the dim.

Generally speaking, engagement algorithms exist for one of three reasons: to sell a user something, to collect user data for later use or sale to a third party, or some combination of the above. A 2018 blog post by The Growth Institute referred to user data as "the new oil" and "arguably the world's most valuable resource" (Ismail). It thus behooves companies to maximize their ability to harvest this priceless commodity wherever possible, and to make user experience a priority so that the flow of data can continue uninterrupted. Companies employ algorithms to great success: YouTube reported in 2020 that 70% of all time spent on the platform was driven by its engagement algorithm (Alfano et al. 2020). Algorithms make use of several key functions to make browsing their platforms more efficient for users, and to make desired content as easily available as possible.

Autocomplete. Search functions, most prominently the ones of Google and YouTube, predict what users want to see as they type and provide easy-access

suggestions. As we will see below, Google's autocomplete function plays a large role in Simpson's (2012) "Evaluating Google as an Epistemic Tool".

Suggestions. These are also commonly referred to as recommendations. When a user first opens up a social media app or website, and thereafter throughout their experience, they will be greeted with content suggested by the platform. What algorithms take as input for suggestions will vary from platform to platform: it will at least be based on a record of the user's browsing habits (i.e. their "cookies", or bits of data compiled by the user's browsing and stored by the website), and may extend to relevant biographical information about the user, especially their location, or past content the user has engaged with by liking it, following the creator, and so on. Depending on the platform and algorithm, the content suggested may be any combination of paid-for advertisements by businesses (most prominently on Facebook), popular content relevant to the user based on things they have liked (common on YouTube), or content liked by other users that the user follows (as on Twitter.)

Filtering. Many parts of a user's experience on a social media platform – including their search results, what the platform advertises to them, and what content is presented to them as they browse – is filtered behind the scenes in accordance with that user's browsing habits and data profile. Hootsuite, a social media post scheduling app favoured by influencers and businesses alike, claims that "[t]he average organic Facebook Page post sees just 0.07% engagement. To bump that up for your brand, you've got to learn how to signal the algorithm" (Newberry 2022). Hootsuite's assertion that brand content will remain largely unseen without the attention of the Facebook algorithm suggests that algorithms have vast amounts of control over what is seen and not seen by

any given user at any given time on their platform.

Technological seduction and epistemic structures

Alfano et al., in "Technologically Scaffolded Atypical Cognition: The Case of the YouTube Recommendation Algorithm" (2020), attempt to glean some useful insights into the way the patented YouTube algorithm works and what role it plays in online radicalization.

The authors make use of two key terms to make their case. They refer first to the concept of technological scaffolding, which broadly understood is the way engagement algorithms help a user find content they may not have been exposed to otherwise. Technological scaffolding plays a role in what the authors refer to as technological seduction, a term borrowed from Alfano, Cheong and Carter (2018), which features the appearance of a discussion between a user and an algorithm as the algorithm engages in "mind-reading" to provide a user with content they didn't know they wanted. To be successful in seduction, the algorithm must assume a position of authority over the user: "The authority in question is epistemic rather than the authority of force. Seduction is distinguished from assault in that it aims at, requires, even fetishizes consent." (4)

To facilitate this "mind-reading" effect, the authors assert that algorithms rely on your "digital footprint", or the data profile collected by companies based on your browsing, searching and purchase history. (15) They assert that more accurate recommendations can be given to you based on seemingly unrelated data. This service is, purportedly, relatively innocuous on its own, engaging the user with content based on their previous viewings and tailoring it to their continued interest, and is at least *prima*

facie beneficial to both user and provider.

One issue with this service, as the authors identify, is the idea that technological scaffolding can take you from supposedly apolitical topics down the rabbit hole of political conspiracy theories. One major takeaway for our purposes is that as this twist of the original intent is not the result of any one individual or group gaming the recommendation system, the responsibility of radicalization seems, at least *prima facie*, to be diffused among algorithm companies and influencers. As the authors assert, this technological scaffolding, and subsequent technological seduction, only has the *potential* to radicalize – it does not appear to be intended to function in just this way (5).

Alfano et al. argue that technological radicalization involves an "iterative, path-dependent process" (11) and that users don't initially receive radicalizing content but get there through "progressive, incremental stages" (ibid.) The authors characterize the presentation of content to the user as happening on several different "levels", and conduct their research by using a web crawler to simulate a user following several "rabbit holes" up to five layers deep (12).

Technological seduction operates on an implied, unspoken discussion between an algorithm (the seducer) and an agent (the seducee) that progresses through several steps of conversation:

The first step is an implicit statement from the algorithm that we can interpret as "I know what you're thinking". The authors refer to this as a "gambit" (4) where the seducer insists they know what the seducee wants, better than the seducee does themself. "'I know what you're thinking' presupposes or establishes an intimate bond. Nothing is more bound up with personal identity than someone's inner life — their thoughts,

feelings, emotions, and values" (ibid.) The algorithm, in the role of seducer, then provides the user with an answer to their question, and the user, in the role of seducee, implicitly responds "Yes, you do know what I'm thinking" by following it.

A successful seduction, then, carries with it the implication of "I know you better than you know yourself." This implication is similar to how Simpson (2012) argues that Google functions as a "surrogate expert". Features of engagement algorithms such as filtered results and the autocomplete function add to the illusion that an engagement algorithm is providing just what you want. But an engagement algorithm cannot be a true expert, only a "surrogate expert" as Simpson argues. Algorithms do not process their recommendations and tailor their results to the needs of the question-asker the way that humans do. They are made to maximize profits, and to guide you to the content that will most likely result in a sale in the short or long term.

In any case, the success or failure of the effect is reliant on a human agent "playing along" with the effect and "acting like" they are actually interacting with an expert (Simpson) or being seduced (Alfano, Carter and Cheong). In a vacuum, this effect is not harmful. However, the more an agent comes to rely on it, the more short-circuited their epistemic pathways become, and the less reliable their testimonial knowledge becomes in turn.

Consider that an algorithm can only do what it was programmed to do, and lacks the ability to make human judgment calls. A physician can use such human qualities as professional expertise, experience, intuition and insight to determine whether a given treatment is appropriate for her patient. A recommendation algorithm, on the other hand, can only bind its recommendations to those values which it has been told to follow by its

programmers. If it has been bound primarily to "usage-maximization", as argued by Miller and Record (2013) and Côté-Bouchard (2020), it will not hesitate before referring a user to increasingly anti-Semitic content, nor will it pause to check his history and make a value judgment about whether more objectionable content would be a good idea. It will merely pick up keywords in the user's history and point him to more content it thinks he might like, based both on his keywords and on how much the new content being recommended has been viewed.[1]

To continue the physician analogy, if an artificial nurse kept prescribing you morphine for your broken arm, based on your past medicating habits and the medicating habits of others in the same predicament as you, that would likely be considered suboptimal medicine. As we will see in the following section, this illustrates precisely *why* engagement algorithms cannot serve as true experts per Simpson's account – because their recommendations of pertinent information are based too much on what they have found that you, the user, find the most engaging and read or watch the most often.

Thus, if human agents turn more frequently to algorithms as knowledge surrogates, and also sounding boards – i.e. places to vent or to voice dark or secret thoughts in hopes of validation, and to receive information that confirms these thoughts, we have cause for concern, as there's no human being on the other end of the conversation to exercise judgment and engage in some sort of intervention process for the agent. This is evidenced in the case of Charleston church shooter Dylann Roof: "According to prosecutors, Roof did not adopt his convictions 'through his personal associations or experiences with white supremacist groups or individuals or others.' Instead, they developed through his own efforts and engagement online." (2)

---

[1] This argument is drawn from my (2021b) manuscript.

Revisiting testimonial networks

Let us briefly reimagine the issue in the framework raised by Sullivan et al. in the previous chapter. Recall that Sullivan et al. were concerned that agents may receive information from a variety of sources that appear to be diverse and, thus, confirm the information reliably and independently, when in fact they all supervene on a single source and are merely acting as signal-boosters. If it is the case that agents are growing increasingly reliant on algorithms as surrogate experts, then the more an algorithm guides an agent towards a certain answer, the more likely that agent is to believe it's true in each subsequent interaction. But, of course, what may be hidden from the agent is that the results provided to her by the algorithm are going to be informed by her data profile and browsing habits – making her, unwittingly, the architect of her own compromised testimonial network.

Algorithms, through their usage of strategies such as filtering, suggestions and autocomplete, are thus likely to embody what Sullivan et al. are concerned about. That is to say, these features will create the illusion of sources "B," "C" and "D" telling an agent that *p* when, in fact, these sources all supervene on "E": the agent's own data profile and browsing habits. Thus, the more reliant an agent becomes on information from a testimonial network that is suboptimal or compromised, the more likely she is to receive suboptimal or compromised information.

Limitations and caveats

Alfano et al.'s research has some methodological limitations. The goal of the

paper is fairly specific: the authors aim only to show that there exists a pathway between regular, innocent topics of interest and conspiracy theories on YouTube, not to guess at whether any agents, or how many, take that pathway (6). By the authors' own admission, it is difficult to replicate what an individual user's browsing preference might be using only a "naive" bot that doesn't have a data profile (37). To some extent, this requires assuming that there are agents out there who are becoming radicalized by clicking mindlessly on the first topics that come up. These agents, presumably, would have to lack any kind of critical thinking whatsoever. While this is possible, it does not adequately capture Alfano et al's conception of "technological seduction", that a radicalization process is a discursive one, or at least a pseudo-discursive one, between an agent and a social media network.

The experiment also does not capture the other ways the process is discursive; reading or participating in a comments section on YouTube, for example, let alone one on Facebook or Twitter, is another way an agent's values and sense of identity are manipulated. In this way, it does not capture the concern we raised when reviewing Sullivan et al., that engaging with content on social media goes beyond being a mindless consumer of content, but also involves the discussion agents have with each other and with themselves. Alfano, Carter and Cheong seem to have this concern in mind when they assert that "We find out what we think by expressing it and hearing it echoed back in a way we can accept; we also find out what we think by having thoughts attributed to us and agreeing with those attributions" (Alfano, Carter and Cheong 4).

Since most humans do not behave like naive bots blindly watching content on YouTube, the methodology used by Alfano et al. creates results that function rather like a

positive control in a scientific study. That is to say, their paper shows one possible scenario if YouTube's recommendation algorithm is the only factor effecting change on an agent's epistemic structure. Nevertheless, Alfano et al.'s study serves as an illuminating one: it provides a very plausible picture of how the YouTube algorithm might accelerate online radicalization with the data available. To properly understand the role played by engagement algorithms in Internet radicalization, however, it may be necessary to understand not just how the algorithm itself works, but how it works *in conjunction* with other components of online radicalization; in other words, how the big picture works, holistically.

Echo chambers and epistemic bubbles

Echo chambers and bubbles are common enough in literature on online radicalization. Alfano, Carter and Cheong note their influence in the process of radicalization, asserting that "[o]n another view (Silber et al. 2007), the Internet facilitates self-radicalization by generating an echo-chamber (AKA "filter bubble"): people interested in radical ideology tend to communicate directly or indirectly only with each other, reinforcing their predilections. A related opinion is that the phenomenon of group polarization, as facilitated by social media chat forums, is among the salient causes (Sunstein 2011; 2017). (2)" However, C. Thi Nguyen, in "Echo Chambers and Epistemic Bubbles" (2020), invites us to think of them a bit differently.

Nguyen argues that while we typically conceive of the two title terms as synonymous, they have very important mechanical distinctions. Epistemic bubbles, he argues, are a kind of non-ideal informational framework where each agent, by design, has

the same views. On Nguyen's account, epistemic bubbles can happen by accident, through ordinary mechanisms of social selection – we tend to agree with our friends, so we want to hear their voices more. These carry some amount of risk to them – as Sullivan et al. argue, receiving affirmations from a plurality of sources can distort one's ability to criticize what one engages with – but are fairly fragile. An epistemic bubble can be "popped", Nguyen thinks, with the proper application of solid contrary evidence.

Echo chambers, by contrast, actively discourage testimony from outside sources. Echo chambers are active manipulations by those in positions of authority of what Nguyen calls chains of trust. A chain of trust, according to Nguyen, is what allows you to trust that a bridge is safe to drive on: you trust the engineer who built it, who trusts the physicist who taught them the principles of their engineering, and so on. When these chains of trust are manipulated, you get something resembling conspiracy theory-style thinking: someone who rejects the canonical picture of belief, such as the chains of trust that allow us to trust in vaccine science, for example.

The concept of "chains of trust" is central to the reliability of testimony, forming a crucial source of exterior information for all agents. Consider the following passage from Coady's "Testimony: A Philosophical Study" (1992):

> "My first morning in Amsterdam I wake uncertain of the time and ring the hotel clerk to discover the hour, accepting the testimony of the voice just as I would accept the institutional testimony of a clock or watch; being early for breakfast I read a paperback history book I have brought with me which contains all manner of factual claims that neither I nor the writer can support by personal observation

> or memory or by deduction from either: the deeds of a man called Napoleon Bonaparte who is supposed to have done all manner of astonishing things more than 150 years ago, many of his exploits being performed in places neither I, nor even perhaps the author, has ever visited and the reality of which is accepted on the word of others [...] No wonder that David Hume, who is one of the few philosophers to discuss the topic seriously, says of testimony, 'there is no species of reasoning more common, more useful, and even necessary to human life, than that which is derived from the testimony of men and the reports of eye-witnesses and spectators'." (Coady 5)

Testimony has featured as a crucial ingredient to form a complete picture of the outside world at least as far back as the middle ages. Aquinas, according to Coady, considers it "natural faith" and "necessary for human society" (13). The existence online of epistemic bubbles and echo chambers serves only to demonstrate Coady's point about the importance of testimony in everyday life – since, in effect, these are also instances of testimonial chains, maladaptive though they may be.

Borrowing from Endre Begby, Nguyen asserts that the application of contrary testimony cannot dislodge someone from an echo chamber; quite the opposite, as those trapped within may be taught to anticipate such evidence, and distrust it when it arrives. This anticipation and distrust of contrary evidence may actually reinforce an agent's trust in the system of beliefs supplied by their echo chamber (Nguyen 12).

<u>The callusing of epistemic bubbles</u>

We might be inclined, on considering Nguyen's points, not to be too concerned about the negative effects of algorithm reliance. After all, much of what we have seen so far may lead us to believe that algorithms play a role in forming epistemic bubbles around agents, and it should be easy enough to dislodge those agents from those bubbles: the right amount of empirical evidence should do the trick. However, as we have seen unfold over the last few years, all too often no amount of contrary evidence can dislodge those who have been radicalized online. This gives us reason to believe that such agents are not trapped in Nguyen's epistemic bubbles, but in fact echo chambers.

What the unfortunate agent trapped in Nguyen's echo chamber appears to exemplarize is "conspiracy theory" style thinking. Conspiracy theory thinking, as I characterize it, is one that denies the foundation of testimony – that is to say, that the person upon whose testimony you are relying is an expert or at least that their view is creditable and reliable. Typically, this takes the form of either denying that the "other side's" foundation is reliable, or otherwise that *anyone's* testimony is valid – so long as it supports your point of view and not the other side's (i.e. "alternative facts"). In either case, conspiracy theory thinking demands that the "chains of trust" that Nguyen describes, and that Coady alludes to, must be subverted. That is to say, it demands that the typical structure upon which we come to find some testimony reliable, but not others, must be rejected. This allows those caught in its snare to come to believe in and trust the testimony of sources most do not consider reliable, such as unaccredited talk show hosts, while rejecting the testimony of traditionally reliable sources, such as scientific experts.[2]

As we touched on previously, the processes of social media filtering over time

[2] This is of course not to provide a comprehensive account of conspiracy theory-style thinking. It merely serves to provide a relevant characterization of how this sort of thinking relates to our topic at hand.

will eventually collapse one's testimonial networks if not monitored and kept in check by the agent. Through technological seduction, an agent grows to put trust in the ability of an algorithm to give her what she wants. Correspondingly, the more an agent puts trust in an algorithm as an epistemic source, the more it learns about what she wants and the better a position it is put in to deliver that to her. This process creates a feedback loop that serves to shrink our epistemic and testimonial structures at both ends, placing us in epistemic bubbles. Over time, and through repeated engagement with and reliance on an algorithm as a surrogate expert, these epistemic bubbles are liable to harden into echo chambers, rather like a blister hardens into a callus.

All of this may or may not be by design, but recall that the overarching goal of an algorithm is to collect data on the user so that it can profit off her, either by selling her a product or by marketing her data. Thus, as far as the company that owns the algorithm is concerned, her continued engagement with and reliance on the algorithm as a source of information proves that it is working as intended. Her dependence on the algorithm, and all of the ugly epistemic consequences that entails, is a feature, not a bug.

Chapter 3: Epistemic Harm

The epistemic objection

So far in the thesis, we have taken as uncontroversial several notions, for example that engagement algorithms play some role in creating epistemic bubbles and echo chambers, that epistemic bubbles and echo chambers play some role in radicalizing individuals, and that the resulting radicalization is normatively bad. When we talk of "normatively bad" consequences, we have in mind something objectively harmful: for example that a radicalized person could do violence to someone, such as an immigrant or other marginalized person who they have been led to believe is threatening their way of life; or perhaps that a radicalized person refuses to wear a mask during a global pandemic, or get an important vaccine that will help protect them and others.

There are, of course, gradations: refusing a vaccine is not the same as committing mass murder in a place of worship, even if by refusing a vaccine you unwittingly contract and pass on an illness that claims as many lives as the act of mass murder. It suffices to say that these are the harmful and "normatively bad" consequences we have in mind, as opposed to trivial ones, such as if people were being radicalized into being fans of the Toronto Maple Leafs and refused to believe that any other team could have a chance at winning the Stanley Cup.

Taking these notions as uncontroversial is important because it allows the focus of this work to be directed at *just what* role is played by algorithms in creating bubbles and echo chambers, *just what* role those bubbles and echo chambers play in radicalization, and *just what* is to be done about these issues.

Recall that in the introduction to this thesis, the Canada Centre for Community

Engagement and Prevention of Violence, a resource I consider archetypal of the popular picture of radicalization mentioned earlier, characterizes radicalization as a phenomenon that occurs on a timeline. According to their primer document, first agents adopt extreme positions (radicalization), then adopt an ideology that justifies the use of violence (radicalization to violence), and finally become proponents of violent extremism, supporting the use of violence to achieve their extreme political goals.

I argue that this picture is inadequate, because it fails to capture what happens when people are radicalized online and participate in disruptive and ignorant group action that does not fall under the popular picture of violent extremism previously characterized, such as the so-called "Freedom Convoy" that occupied Ottawa from late January to late February of 2022. In order to capture the actions of those occupiers as an example of radicalization, we must accomplish one of two things: we can argue that this radicalization is still bad and harmful even though it doesn't involve the popular picture of violent extremism previously characterized; otherwise, we can argue to change that picture itself. Of the two approaches, we will pursue the first; I believe it is, in fact, necessary to change the previously characterized picture of violent extremism, but doing so is largely outside the scope of this thesis.

Why I aim to problematize the "pipeline" or "timeline" picture of radicalization is that people don't have to commit acts of terror and mass violence to be considered harmed by the effects of online radicalization. Many participants in the "Freedom Convoy" were victims of online radicalization and misinformation, but did not intend to commit mass violence. It suffices to say that these agents and others like them are "epistemically" harmed, that their means of gathering information upon which to base

their beliefs has been compromised in some way.

The epistemic harm argument

"Active harm" or being radicalized to violence is, thus, neither a sufficient nor necessary condition of Internet radicalization. The actual harm we should concern ourselves with is to the epistemic capacities of the agent: how social media algorithms are reshaping and compromising our abilities to form justified belief based on reliable testimony.

Suppose someone, H, finds themselves in an echo chamber where they are led to believe that they ought to devote all of their time, resources and energy towards fundraising to cancer research. Recall that echo chambers, per Nguyen, are different from epistemic bubbles in that someone trapped in an epistemic bubble can be easily rescued from one with the presentation of solid contrary evidence. Echo chambers, by contrast, demand from those trapped within them the mistrust of *any* contrary evidence, and *any* source that seeks to provide same. If H is trapped in an echo chamber, it is plausible that they are led to believe that dedicating all of one's time, resources and energy towards fighting cancer is not just a good thing, but *the only worthwhile thing worth pursuing*. H is thus led away from other normatively good pursuits, such as volunteering in soup kitchens, or even from their own aspirations, for the sake of dedicating their life to fundraising for cancer research.

Now, if I understand this concern correctly, then as alarming as this opportunity cost is, it is really only an unfortunate consequence of the real problem: H's epistemic enslavement to an echo chamber, starving their knowledge and belief heuristics and

rendering them a poorer, less capable agent. Call this the *epistemic harm argument*: that the real damage of radicalization is the ways in which it hamstrings an agent caught within from gaining knowledge and forming beliefs in ways that are acceptable to us. The epistemic harm argument holds that normatively bad consequences, such as someone being driven to violence as a result of their radicalization, is neither a sufficient nor a necessary condition of defining why radicalization itself is bad.

Epistemically harmed agents will be more vulnerable to many of the phenomena we have already observed. Their testimonial capacities will be shrunken and vulnerable in the way that Sullivan et al. describe, and they may receive a great deal of signal boosting stacked precariously on a vanishingly small amount of real sources, constrained further still by their own online habits. They may be more easily swayed by the mechanisms of technological seduction. All of these things make them more likely to believe misinformation. They may not be aware of some, or any, of these vulnerabilities. Finally, and most damningly, because their online browsing habits both reflect and shape the picture of what they see online, the more epistemically harmed an agent is, the more harm they are likely to incur.

Miranda Fricker, in *Epistemic Injustice: Power and the Ethics of Knowing* (2007), cleaves between two main kinds of epistemic injustice, which she calls testimonial and hermeneutical injustice. Both of these wrongs, Fricker argues, are unjust because they wrong us in our capacities as knowers, which are capacities essential to human value (4). Testimonial injustice, argues Fricker, occurs when an agent's report is not taken seriously because of some prejudice against them. For our purposes, we will be more interested in hermeneutical injustice.

On Fricker's account, a wrong is hermeneutically unjust when the experience of the wrong is poorly understood by the victim and those around them; the frameworks and linguistic concepts to parse and discuss the wrong are underdeveloped.[3] This seems to fit well with agents suffering epistemic harm as a result of exposure to online engagement algorithms.

First, many agents lack the vocabulary and understanding of computers, cell phones, the Internet, and algorithms to understand why their data is important, why it might be collated in a data profile, why it might be in their best interests to try and protect that data profile, and just what they can do to protect themselves online. Second, many agents lack the vocabulary and understanding of epistemic harm itself; how being epistemically harmed in the way I describe can cause an agent to act, and why they may be unable to identify it, avoid it, or indeed keep themselves from continuing to seek it out.

Fricker argues that these kinds of injustice are structural, as a result of a mismatch of either material power or identity power, or some mixture of both. To wit, Fricker argues that no one individual is necessarily culpable as "harmer" in order for harm to be identified. This makes her framework an important one for our purposes, since we will see in the next chapter why social media companies share at least partial responsibility for online radicalization simply for having created an environment that can cause epistemic harm, even if that is not the intent of the platform.

---

[3] As Fricker has certain experiences in mind, such as sexual harassment or racialized violence and other wrongs experienced by marginalized people, hermeneutical injustice is not a strictly appropriate term to apply to Internet radicalization. The concept itself, however, is relevant and helpful in characterizing the epistemic effects of Internet radicalization on an individual.

Caveats

Is this altogether satisfying to our argument? Even if we determine that it is the mechanism of radicalization itself that is normatively bad, have we said all there is to say about the kinds of radicalization happening online? It doesn't seem inconsequential to our argument that certain people are being radicalized towards violent, hateful movements, such as the Proud Boys and QAnon. Or, as a corollary, that there are no stories of people becoming radicalized online to spend all of their time volunteering in soup kitchens, forming groups and movements around soup kitchen action and actively resisting evidence that soup kitchens may not be the most beneficial way to spend their time and goodwill. Is it just a coincidence that much of online radicalization has certain kinds of beliefs in common? Worse, is this observation just the cognitive bias of the researchers investigating this problem?

Perhaps we might amend the epistemic harm argument by arguing that if radicalization itself is epistemically harmful, then it is even worse if agents are being led into epistemically harmful environments where they are led to believe harmful or violent things about other people. We don't want to argue, after all, that radicalization *just is* this epistemic damage; someone trapped in an epistemic bubble or echo chamber, such as our hapless Leafs fan, might be brainwashed, but it would be a stretch to say they were radicalized. The *character* of the radicalization, or just what makes up the content of the views an agent is subjected to, seems to matter.

And, conversely, we don't need Internet radicalization to create identity-based violence. A Manchester United fan may incite violence in a pub against a fan of Liverpool, but it seems strange to call him "radicalized" as we understand it in the online

sense. To borrow Nguyen's distinction, he doesn't seem to be caught in an epistemic bubble; he may have friends and family that like other soccer teams. Nor does it seem right to say he is caught in an echo chamber, as the Manchester United fan club does not appear to be an abusive cult that discourages its members from listening to anyone who says that Man U is not the best club in the world. In fact, his inclination towards violence doesn't seem to have much to do with his epistemic state at all. He does not appear to be either radicalized or brainwashed.

Paul Gow and Joel Rookwood, in a 2008 study, found that "involvement in football violence can be explained in relation to a number of factors, relating to interaction, identity, legitimacy and power." (Gow and Rookwood 71) Paul Katsafanas, in "Fanaticism and Sacred Values" (2019), seems to agree, characterizing fanatics as in need of group identity and ensconcing certain values as sacred. Katsafanas asserts that characterizing fanaticism in this way is necessary to correctly critique it as promoting a certain kind of social pathology, as opposed to arguing that it rests on incorrect values or beliefs.

Following Katsafanas, the underpinnings of fanatic behaviour – to wit, a search for group identity and shared values – have much in common with the mechanics of online radicalization. Thus, I think we ought to characterize online radicalization not as an entirely new phenomenon, but rather as a mechanism of accelerating the radicalizing means we already know of, only in unpredictable and exponential ways. We can thus conclude that radicalization is not just a buzzword, or worse, a red herring disguising the real factors behind group violence. Rather, as we characterized it in the introduction, radicalization is a discursive process that changes the viewpoints of an agent to one that

is extreme in a number of ways. This indeed seems to be one function of Alfano, Carter and Cheong's (2018) concept of technological seduction: a discursive, iterative process that initiates an agent into certain viewpoints they might not otherwise have been exposed to quite as directly.

In an article for Newsweek, Jeffery Martin reported that Alexis Ohanian, co-founder of Reddit, believes that social media helps to create an environment where conspiracy theories seem more plausible to the average user. According to Ohanian, "[t]here is a legitimacy that comes with seeing a hate group or conspiracy theory in a feed right alongside your uncle celebrating his promotion or some cute photos of your nephew." (Martin, 2020) Ohanian went on to assert that certain individuals sign onto believing conspiracy theories because it helps them "feel normalized and feel kinship and community around them." (ibid.)

Online radicalization, then, while posing serious epistemic risks for those caught in it and thus being *pro tanto* normatively bad per the epistemic harm argument, also appears to have a relationship with the kinds of factors – namely, a search for shared values and group identity – that can dispose one to become a soccer hooligan. It could be argued that one role played by algorithms is to *amplify* these factors: someone, S, could be more susceptible to enacting violence on a marginalized person because S feels S's identity is at risk, a feeling that is constantly reinforced by S's online communities.

I think, however, that we should be more concerned with the ability of algorithms to *accelerate* epistemic harm: that interacting with engagement algorithms risks a seductive feedback loop poorly understood by the agent – perhaps not even by the company employing the algorithm. Once caught in this feedback loop, agents grow more

likely to rely on algorithms for information, and simultaneously more susceptible to subsequent epistemic harm; rather like how biting a canker sore will inflame it and cause repeated biting, more use begets more harm, begets more use, begets more harm.

By design

Recall that engagement algorithms are typically money-making devices constructed by social media companies to do one of several things: encourage the engaged agent to make a purchase, build a data profile about the engaged agent's interests to sell to advertisers or to drive internal advertising, or some combination of the above. To this end, it does not behoove social media companies to keep tabs on *just what* a user is viewing, or make qualitative value judgments about the health, safety or credibility of a user's browsing habits, because they are value neutral except where it concerns the generation of profit. That isn't to say that companies don't analyze their own data, but their analysis is typically focused on the goal of maximizing profit, not on the goal of rooting out and eradicating misinformation. In effect, all that social media companies are doing is erecting a vast and efficient infrastructure – we can reimagine the garish term "information superhighway" – for the purpose of becoming players on the data market.

This infrastructure, overseen and optimized with the overarching aim of profit growth, is susceptible to the formation of epistemic bubbles and echo chambers by content creators. Many of these content creators are merely misguided and relatively benign, amplifying the voices of discreditable sources because it suits their viewpoint. However, some are actively malicious, and aim to recruit users to groups where

"conspiracy theory" style thinking reigns, such as hate groups, anti-vaccine groups, flat Earthers, and so on. The issue isn't to determine who is to blame: to try to cleave the benign content creators who repost misinformation and reinforce an epistemic bubble, and the malicious ones who hijack the algorithms to create echo chambers would be to "chase ghosts", so to speak.

Keeping the epistemic objection in mind, then, what is especially bad about online radicalization is not just that it epistemically hamstrings ensnared agents, but that it does so in a way that is especially accelerated and unpredictable – and nearly impossible to grasp, track, check or stop. Online radicalization and its associated epistemic harm is accelerated by the infrastructure of social media algorithms: this is the tacit role played by social media companies in the radicalization of online users.

Chapter 4: Who Is Responsible?

Why algorithm companies are responsible for epistemic harm

This will be the shortest section of the thesis, as the moves within follow quite logically, I think, from the implications of previous sections.

By now, it is hopefully clear to the reader that engagement algorithms constitute an environment that can cause epistemic harm to the capacities of the agents who use them. To recap, I identify three main processes by which engagement algorithms cause epistemic harm:

i. technological seduction: Alfano, Carter and Cheong's concern that algorithms engage in a pseudo-discursive process with agents whereby they purport to know what agents are thinking, "finish their thoughts" for them, and provide them with content they didn't know they wanted;

ii. structure-shrinking: Sullivan et al.'s concern that an agent can be exposed to a variety of views that seem independent, thus providing testimonial accuracy to a viewpoint, when in reality those views are not independent at all and serve to signal-boost the viewpoint of a single source;

iii. unconscious self-guidance: an extension of Sullivan et al.'s concern, when the seemingly independent signal-boosting voices are manipulated by a user's data profile and supposed preferences, and the single viewpoint they supervene on consists in the user's data profile.

In "Evaluating Google As An Epistemic Tool" (2012), Thomas Simpson argues that users require search engines to play two roles: *navigational* roles and *informational* roles. A user relies on a search engine to fulfill an informational role when looking for a

general answer to something they may not know very much about already. By contrast, a user's preconceived notions and prior knowledge will inform the way they use the search engine in a navigational role. Users require search engines to operate navigationally when looking for specific content to fill in a knowledge gap (e.g. attributing a quote) and, thus, the algorithm nominally cannot be entirely responsible for the results, since it operates more like a "searchlight".

For the purposes of this section, I aim to engage primarily with Simpson's argument, which focuses on the efficacy of Google versus other search engines. Though a search engine still features in every social media platform, and is still a primary way of interacting with and navigating services such as Google or Amazon, it should be noted that some newer platforms, like TikTok, aim to provide a heavily curated "newsfeed" service to the user that facilitates endless scrolling without the need for user queries, so that the search engine is not much more than an afterthought for the purpose of most users' engagement. Nevertheless, much of what follows can be expanded to encapsulate most any kind of user navigation through a social media platform.

It should not be very controversial that engagement algorithms appear to manipulate the informational role played by search engines. By now, we have established that algorithms aim to prioritize content they think the user is more likely to engage with, so that two different users with two different profiles searching for the same content are likely to be presented with different results, or at the very least differently-ordered results. To this end, Simpson argues that search engines should be evaluated for their epistemic qualities based on their timeliness (the amount of time it takes for a user to find a result she's looking for), authority promotion (ability to prioritize truthful testimony

from experts) and objectivity (ability to equally prioritize all relevant views on a topic).

Simpson argues that "[f]or all the criteria of assessment except objectivity, there is an alignment between the interests of search engine operators, the preferences of the private individuals who use the service, and publicly desirable outcomes (440)." The timeliness and precision of a search engine is ever the goal of the company in charge of an algorithm, since if a search engine does not provide timely relevant results, it is unlikely to be popular, and thus succeed in its goal of maximizing profit. In this way, a company's goals align with that of its customers and of the public good for many other criteria, including authority promotion and two others posited by Goldman, precision and recall. However, I do note that while search engines do aim at authority promotion, the data profiles of some users may indicate that they are not interested in truthful testimony from experts, and so those results may not be prioritized for those users.

Objectivity, Simpson argues, is different. "Descriptively," he argues, "and pessimistically, most people do not care terribly much about objectivity," and so there is little commercial incentive for companies to prioritize this criterion (441). Speaking conservatively, search engines are "objective", so far as Google is not nefariously injecting their own perspective and the political perspectives they want to see their users take into search results. But so long as search results are shaped by a user's data profile, they are not going to be objective.

While it would be too strong to say that corporations are *directly* responsible for how people create content on their platform, I contend that they are *indirectly* responsible, insofar as much of what users see on a given social media platform is tailored (i.e. filtered) to what the platform's algorithm thinks they want to receive. This,

in turn, will influence how users respond with their own content. Since corporations tightly regulate and control how we access (and therefore use) their platforms, I argue they have a *de facto* responsibility for how people create (and engage with) content on their platform.[4]

In addition to manipulating the informational role, I argue that engagement algorithms are also responsible for manipulating the *navigational* role played by search engines. In other words, exposure to engagement algorithms online, over time, will change the sort of things that users search for. The content, influencers, groups, and language that users are exposed to will change the information they look for, the words they use to look for it, and the results they embrace – or eschew.

Consider if a curious agent, A, wondered about methods to improve her health, energy and diet, and had heard good things from her friends online about multivitamins. A goes on to Google "benefits of multivitamins", and likely receives search results extolling the many virtues of multivitamins, including a great deal of "infotising" articles (e.g. Masnick 2003).[5] On the other hand, agent B also looks to improve her health, energy and diet, but is skeptical about multivitamins, since she has heard online anecdotes that claim they are useless, or even harmful. B goes on to Google "are multivitamins useful", and likely finds search results doubting the usefulness of multivitamins in a healthy and balanced diet, perhaps even cautioning the user against taking them in lieu of a proper diet. In both cases, the way A and B phrased their queries was influenced by the information they had previously received from online friend

---

[4] This argument is drawn from my (2021b) manuscript.

[5] These typically take the form of advertisements veiled as blog posts that explain a problem and detail a variety of ways to solve it, always leaving the host company's product as the best solution for last. In fact, "infotising" articles likely make up a good amount of positive and negative arguments for a given product, lifestyle choice, diet, exercise, etc. on the Internet, but to assess the depths to which they have muddied the waters is outside the scope of this work.

groups and, perhaps unconsciously, they directed their searches to receive the information they thought they wanted.[6]

This is all seemingly anodyne enough, but the way users make use of search engines in recent times has contributed significantly to the spread of misinformation. Consider, now, that A is worried about her health and safety, and the health and safety of others around her, during the COVID-19 pandemic, and is wondering if recently rolled-out vaccines are able to protect her, her loved ones, and people she interacts with in her day-to-day life. She Googles "covid vaccine effective", and is likely to receive positive results assuring her of the safety and efficacy of vaccination against COVID-19. Meanwhile, B is skeptical of vaccination, since she has heard online that vaccines have dangerous ingredients, or are vehicles for government experimentation. She searches for "risks of covid vaccine", and is liable to find some misinformation baselessly warning her about the harms of vaccination. Note that both users' predispositions, habits and prior beliefs will all work in holistic tandem with one another throughout this process, since these things influence what a user has searched for before, and what their data profile looks like now. This means that B is *especially* likely to receive these results if her data profile suggests that she wants this misinformation.

Expanding upon the characterization of "unconscious self-guidance" above, engagement algorithms play a role in altering how epistemically harmed agents search for content in a navigational way. In both navigational and informational uses, a user's data profile will reorder and modify her search results, prioritizing content the algorithm thinks she is more likely to engage with, since it is sympathetic to her viewpoint.

---

[6] Let the skeptical reader be reminded that there is nothing magical about online discourse, and that A and B could just as easily have been motivated to phrase their queries based on things they had heard in person. What is special about online discourse is not its supernatural convincing power, but its ability to be pervasive, constant and unconsciously manipulative.

Additionally, an epistemically harmed agent is more likely to word her query in such a way, consciously or unconsciously, to receive the information she wants. Instead of "blm protests seattle", she is more likely to search for "antifa riots seattle"; instead of "covid vaccine effective", she is more likely to search for "5g vaccine chip". Since, when she gets the results she wants, she engages in a successful technological seduction, this creates a powerful feedback loop.

The nature of epistemic harm

As I use it, "epistemic harm" is not some sort of nebulous or theoretical phrase that cautions some vague warning of harm in a deontological sense. Epistemically harmed agents act differently than agents without epistemic harm. To our point, I also think they act differently, and less predictably, than agents harmed in physical, or hedonic, ways. Note that this is likely for all sorts of harm and pain that is not hedonic: harm to one's reputation, or harm to one's self-esteem, for example, are also likely to present in unpredictable ways compared to hedonic harm and pain. However, since we are talking about being harmed in our capacities as knowers and thinkers, it stands to reason that we must expect actions and habits that seem counter-intuitive, to say the least.

Compare someone with an optimally functioning body to someone with a significant back or knee injury. The injured person will complain of pain, their mobility will be reduced, they will be unable to lift heavy things, they may need to sit or lie down for long periods of time, and so on.

Epistemic harm is different and can present less predictably. Epistemically harmed agents will not complain of "epistemic pain"; in fact, it is likely they are unable

to report any epistemic damage at all.[7] They won't think more slowly than non-harmed agents, either. But it seems right to say that their ability to do the "heavy lifting" of belief is compromised in a way that seems oblique at first: for example, they believe stranger and stranger things. They become less concerned with whether something is truthful, instead becoming more inclined to believe things that have the quality of "truthiness". Recall from the introduction that things with the quality of truthiness are more concerned with "feeling" true, appealing to an agent's perceptions or intuitions, rather than being logically sound or representing something factual or evidential. Because these agents are likely to fall into these epistemic potholes along the way of their belief formation, their capacities for acquiring genuine knowledge are hampered.

Thus, the more misinformation an agent is seduced by, the more likely they are to strike out in search of more misinformation (requiring a navigational role), and the more likely they are to receive it (as the algorithm orders content according to their data profile and supposed preferences). Since, per Fricker, epistemically harmed agents have their *capacities* hampered, I argue that epistemically harmed agents are not completely responsible for the way they operate search engines in a navigational way. The mechanism that caused them epistemic harm must be at least partially to blame; to wit, the engagement algorithm that seduced them.

The role played by algorithm companies

To illustrate why algorithm companies hold at least a partial responsibility for the epistemic harm and radicalization that occurs on their platforms, consider the earlier

[7] Some canny agents might report "brain fog" or difficulty concentrating, but I think those who can report such symptoms are unlikely candidates for radicalization.

argument that engagement algorithms control many parts of a user's interaction with an app. While a user can search for whatever she wants on a social media app, her search results will be modified and reordered by whatever the algorithm thinks she wants to see, drawing from her data profile or "digital footprint" (Alfano, Carter and Cheong 2018). It therefore may not be sufficient for social media corporations to put the onus on users to protect themselves from misinformation and radicalization while using their platforms, since, even if only accidentally, their algorithm significantly shapes the way the end-user interacts with their site.[8]

The value of "usage-maximization", broadly understood as the aim of optimizing a user's time spent on a service, is an overarching value controlling the collection and usage of user data by companies on the accounts of Miller & Record, Simpson, and Côté-Bouchard (2020). Simpson and Côté-Bouchard both argue that corporations have a strong incentive to encourage usage-maximization in their user base, since usage-maximization is a primary driver of profit. To encourage usage-maximization is to optimize the visibility of attention-grabbing content that is sympathetic to a user's beliefs, typically at the expense of content that is less engaging or less sympathetic. (7)

Thus, if it is the case that users are becoming radicalized as a result of engaging with content on social media platforms, and the platform's algorithm directly controls how much radicalizing content a given user is subjected to, *and* the user does not have full freedom over the way she interacts with website content, then the corporation is morally responsible for their radicalization, even if that radicalization is an unintended consequence of developing a purely profit-seeking algorithm made to maximize user time and engagement on a platform.

---

[8] This argument is drawn from my (2021b) manuscript.

Of course, corporations can never be *completely* responsible for the radicalization of users on their platforms, just like they can never be *completely* responsible for the actions of agents, the way users search for content, or the way users choose to interact with their platforms. But they *can* be responsible for creating an environment that causes epistemic harm to the capacities of an agent, and thus an environment that *does* play a role in radicalization, even if that is not the intended purpose of the environment.

Chapter 5: What Is To Be Done?

The case for and against vigilance

Thus far in the literature, most authors advocate conservatively for a certain set of epistemic virtues practiced by the end-user. Miller and Record (2013) argue that individual users ought to embrace a set of epistemic virtues to ensure that what they read and share online constitutes justified belief.

Alfano et al. (2020) suggest a number of approaches the individual might take to avoid the epistemically detrimental effects of social media. Chief among these are altering several common social media habits, such as turning off the autoplay button, and practicing what the authors refer to as "epistemic vigilance", identifying virtues such as curiosity, humility, attentiveness and open-mindedness as virtues to maximize in the face of an uncertain epistemic environment. Simpson (2012) similarly laments that only "epistemic saints" are safe from the effects of engagement algorithms and recommends virtuous behaviour changes with the hope of prioritizing epistemic values.

One drawback of many of the approaches so far stated is a reliance on virtue ethics. Consider the "epistemic vigilance" advocated for by Alfano et al. This is all well and good – but an unfortunate implication of the position is that people are captured in echo chambers because they are arrogant, intellectually lazy, dismissive and closed-minded. Perhaps possessing the negative qualities listed will make someone more likely to be radicalized, but we have good reason to believe possessing the right virtues is no guarantor of epistemic virtuosity.

Perhaps people become racist and radicalized for a lack of virtue, but I have argued that lack of virtue needs not play a role in their radicalization. It's certainly

possible, indeed plausible, that some agents are more inclined towards radicalization because they are naturally "epistemically lazy", and intrinsically form their beliefs based on things that are convenient and "feel" correct. However, in the previous section and throughout, I have argued for downplaying the relevance of any intrinsic "epistemic vice" in an agent, since the harm caused to the epistemic capabilities of an agent by engagement algorithms makes any individual agent's "intrinsic vice" difficult to measure and irrelevant besides. In any case, any appeal to "epistemic virtue" will almost always happen in a forum where those who are sympathetic to the claim already possess epistemic virtue, and so do not need the appeal, and those who are not sympathetic to the claim probably do not possess epistemic virtue, and so are likely to dismiss it.

Other authors, such as Nguyen, think we should not be so quick to put responsibility solely in the hands of the user. According to Nguyen, "echo chambers prey on our epistemic interdependence. Thus, in some circumstances, echo chamber members do not have full epistemic responsibility for their beliefs. Once one is trapped in an echo chamber, one might follow good epistemic practices and still be led further astray. And some people can be trapped in echo chambers because of circumstances beyond their control — for example, they can be raised in them." (Nguyen 4)

Further compounding the problem, Miller and Record point out that algorithms are "opaque" by design, protected trade secrets that do not reveal their workings to their audience. This is another way in which we might understand this problem as epistemic injustice. To wit, a flaw in virtue-based normative arguments is that, *by design*, not every user knows exactly what they are doing online or how what they do works; advocating for better conduct from those users is difficult when they are kept in the dark about why

they should do those things, or especially if it is easier not to.

Even young people who have grown up with social media don't always fully understand how engagement algorithms work or how to best practice data privacy, usage limits and other "virtuous" aims. Consider, then, the older user who was not exposed to social media until some phase of adulthood. They may not fully understand what data is, or what their data profile consists of; they may not even be confident in operations of a computer that younger people would consider "basic", let alone how to turn off the autoplay button on YouTube, and understanding why they should do that besides.

If the problem really is in the end-user, it may be quite a bit more difficult to solve. Nguyen proposes a "radical rebooting" of an individual's belief system, which relies on mending the manipulations of the chains of trust. He points to an example of a Jewish student who invited a violent neo-Nazi to Shabbat dinners repeatedly and slowly earned the latter's trust, eventually leading him to repent and abandon Nazism (Nguyen 33). However, I think we ought to be wary about any solution that requires a vulnerable person to be responsible for re-educating the person threatening them.

The case for and against legislation

Recall that an important subset of Fricker's (2007) concept of epistemic injustice is hermeneutical injustice. Hermeneutically unjust wrongs are poorly understood by the victim and those around them and occur when the frameworks and linguistic concepts to parse and discuss the wrong are underdeveloped. One way to combat hermeneutical injustice as it relates to online radicalization would be to demand transparency from algorithm companies.

In addition to their stance on epistemic virtues, Miller and Record argue for more transparency from algorithm companies. Some companies have indeed become more transparent about how data is collected and used, and given users a greater degree of freedom to opt out from certain processes. However, there is rarely any certainty how much control users have over their own data, or how thorough or permanent these opt-out processes are; it is hard to say if this is proper choice, or just the choice to rearrange the deck chairs on the Titanic, so to speak.

Part of the worry of radicalization is that a lack of transparency from algorithm companies may cause epistemic harm to an agent who makes use of a system neither they nor others around them properly understand. And it is the case that no one properly understands the short- and long-term harm potential of engagement algorithms on agents: not governing bodies, nor academics, nor algorithm companies, nor influencers, nor users. If an agent trapped in an Internet echo chamber is caught there by forces they do not fully understand, they are *vulnerable*. No discussion of epistemic virtue nor presenting them with the transparency options or "opt-outs" available to them is likely to rescue them from the echo chamber. Internet echo chambers seem to exist *in spite of* a remarkable amount of contrary viewpoints and, in at least some cases, contrary evidence; despite the existence of things we might take to deflate an agent's radicalized position, they often remain stubbornly locked in place, consonantly with Nguyen's definition of echo chambers, rather than epistemic bubbles. It is thus key to my argument that we recognize the vulnerability of such agents and *treat them* as vulnerable. This entails advocating for a more radical legislative approach rather than appealing to a set of virtues likely only possessed by those of us not already trapped in Internet echo chambers.

Megan Fritts and Frank Cabrera (draft) share the concern that agents' epistemic capacities are being tainted in ways that are not well understood. In their paper "Fake News and Epistemic Vice: Combating a Uniquely Noxious Market", the authors argue that there is a market for fake news in two ways: directly, in that fake news websites such as the *Denver Guardian* sell advertisement spots to wealthy corporations such as Google; and indirectly, in that users of social media platforms sign away their data for use in targeted advertising.

Fritts and Cabrera argue that we can understand this market for fake news as a noxious market, per Debra Satz (2010). Satz defines four dimensions along which a market can be noxious:

1. Individual Harms – the market produces harmful outcomes for individual participants

2. Societal Harms – the market is harmful to the society in which it operates

3. Weak Agency – the market involves or requires weak or asymmetrical agency on the part of its participants, and

4. Vulnerability – the market reveals the vulnerability of one of its participants.

Satz asserts that the harms of most every market can be tracked along one of these dimensions, but notes that there are several reasons why we may not want to legislate against these markets. Some markets, however, are too intrinsically harmful and their harms cannot be mitigated; Satz has in mind, for example, organ-selling and sex-work markets.

Fritts and Cabrera argue that the market for fake news is noxious on account of individual and societal harms. In particular, they point to the Pizzagate incident as an

example of societal harm posed by the market for fake news (Fritts and Cabrera 6). "Pizzagate" is a conspiracy theory that in the emails on Hillary Clinton's computer there exists a connection to a child sex trafficking ring to Comet Ping Pong, a pizza restaurant in Washington, D.C. In 2016, a lone gunman stormed Comet Ping Pong with an AR-15 with the intent of killing members of that ring.

Per the epistemic objection, we ought not to overly concern ourselves with incidents such as Pizzagate; if we wish to apply Satz' noxious markets framework, then it is better to understand the individual harm as the epistemic damage done to an agent by their radicalization, and the societal harm as the same on a group scale. But again, we should respond that if the epistemic objection is chiefly concerned with epistemic harm, it also serves us to investigate just what the character of that epistemic harm is; what people caught in echo chambers are being told, and what they might be inclined to do as a result.

Crucially for my argument, Fritts and Cabrera identify the flaws in the "epistemic virtue" arguments proposed by authors such as Miller and Record (2013) and Simpson (2012). Rather than falling prey to some epistemic vice, such as laziness or stubbornness, Fritts and Cabrera argue that it is simply that the background situations of agents, inundated with fake news as they are, have polluted our confirmation heuristics; to the point, they go on to argue, that any kind of education about critical thinking, or teaching of epistemic virtue, is not likely to fix the problem (13).

The authors conclude by suggesting a legislative solution to the problem: amending or repealing Section 230 of the Communications Decency Act, with the goal of holding site owners accountable for damages caused by the content on their sites (21).

They express concerns that this solution could be overbroad, and that the damages concerned are too indirect to properly hold site owners liable (ibid.) My concern, similarly, is that data monoliths such as Facebook and Alphabet are quickly becoming too large to be bothered by legislation, and a more severe solution, such as an antitrust suit that breaks up the megacorporations that own our data, may be necessary.

Many of these accounts have in common a prescriptive account focused on the end-user, but say little about "top-down" prescriptions, or what can be done at the level of the algorithm company. We can identify a few plain reasons for this: perhaps most obviously, philosophers are not typically policy analysts or legislators, and thus it is easier for them to limit their prescriptions to the individual. But other authors give us good reason to advocate for this cautious and conservative approach as well.

In "The Fight Against Doubt", Inmaculada de Melo-Martín and Kristen Intemann identify what they call normatively inappropriate dissent, or NID: "dissent that fails to advance or hinders rather than promotes the aims of knowledge production" (de Melo-Martín and Intemann 145). The authors think it's clear that NID exists, pointing to examples of it in climate change denial and anti-vaccine movements for example. However, they think that attempts to characterize *just what* NID is will capture examples of dissent that is appropriate, productive or helpful. Attempting to characterize NID, simply put, will yield false positives, as likely to catch the sort of dissent that has pushed scientific development along as it is to catch the sort of dissent that hinders it. Further, the authors argue, top-down legislative attempts to crack down on NID can backfire; those who hold NID-fueled views, like the radicalized agents in my account, can get fired up perceiving views they hold being "muzzled" or "censored" by the government.

The authors are therefore not particularly enthusiastic about focusing on NID, since they think the benefits of appropriate dissent outweigh the harm that inappropriate dissent can cause. Instead, they advocate for a "reaching-across-the-table" approach. Those who are resistant to accepting the consensus on climate science, they say, may need convincing that climate change will affect them directly, or that the benefits of climate regulations will outweigh the costs (148). In other words, the authors assert, many instances where people promote or subscribe to inappropriate dissent can be traced to an underlying set of values that cause people to protect their viewpoint at the cost of accepting new information. In order to reach these people, they say, some effort must be made to understand the underlying values at play, and to appeal to those values. This echoes the findings of writers such as Katsafanas.

Properly understood, I believe de Melo-Martín and Intemann's argument represents the strongest opposition to a top-down legislative approach. de Melo-Martín and Intemann build on work that criticizes the "value-free ideal", a position held by some scientists and philosophers of science who state that the role of societal values should be minimized when analyzing scientific theories and evidence. Douglas (2009) has problematized the value-free ideal, asserting that values of any stripe – background, newly formed, conscious or unconscious – inform the work of professionals, even scientists: the biases they approach their work with, what they choose to focus on as they collect and collate data, how they interpret their findings, and so on.

So too can we understand the inner workings of algorithm companies. Miller and Record assert that the only value driving algorithms is "usage-maximization" – more engagement with the algorithm is thought to maximize profit, and make the algorithm a

better driver of profit over time. Certainly, however, we can see how other background values may influence this process as well. Ought we to claim that algorithm companies should practice the same epistemic virtue we demand of the end-user?

The case for and against mediation

Throughout all this, we have the autonomy of the corporation to consider: how to impose regulations upon a corporation's actions, and fathoming what sort of responsibilities it can be expected to have in this matter. In my (2021b) manuscript I argue that corporations must first own their role and acknowledge the lack of robust values guiding their practice before adopting any new set of values. Another difficulty in this process is determining how to enforce such a change, and who would be responsible for doing so. If we are not in favour of government legislation, perhaps an approach of public mediation might also meet with success.

I argue that public mediation is preferable to letting companies sort it out for themselves, since we have good reason to believe companies have been less than successful in internal mediation attempts to combat radicalization. For example, The Wall Street Journal's "The Facebook Files" found that Joel Kaplan, vice president of global public policy for Facebook, has been responsible for stonewalling attempts by internal Facebook teams to promote social good, such as reducing the influence of hyper-partisan "super-sharers", or a project called "Common Ground" that aimed to bring politically divided users together over shared interests such as hobbies. Kaplan reportedly shut down these proposals for fear that right-wing pages, responsible for driving up

engagement with Facebook to a large extent, would be disproportionately affected. (Statt)

Waheed Hussain and Jeffrey Moriarty, in "Accountable to Whom? Rethinking the Role of Corporations in Political CSR" (2016), claim that the only organizations that should represent a group of citizens in social deliberation are politically representative organizations (PROs) – political parties and NGOs, for example. Corporations don't count as PROs on this view, since people affiliate themselves with corporations primarily for economic reasons, i.e. to make money, and not for social ones. Thus, on Hussain and Moriarty's account, these corporations cannot represent citizens, and cannot make decisions in the social deliberations that involve them.

In my (2021b) manuscript I contend, in line with Hussain and Moriarty, that while corporations are undoubtedly stakeholders in the field of human data, they can't really be trusted to be decision-makers in this field. It would be wrong to say that corporations who deal in data don't deserve a seat at the discussion table; on the contrary, they ought to be there to put their cards on that proverbial table! Still, their role would ideally be limited to showing how their algorithms operate, and during deliberation, ultimately agreeing to be bound by the standards placed upon them by PROs.

Prescribing a solution

The most glaring difficulty in demanding corporations mitigate or abandon the value of usage-maximization is simply: with what values should we replace it? Simpson (2012) develops an important framework upon which others might build: recall that he argues that search engines should be evaluated for their epistemic qualities based on their timeliness, authority promotion and objectivity. Simpson also argues for limited public

intervention, on the grounds that objectivity is not a goal that tech companies have a commercial incentive to follow, but it *is* a public good; simply requiring Google to set their default settings from personalized to non-personalized search may be enough to fulfill this goal, he thinks. (441)

Such other values are really *epistemic* values rather than, say, financial ones, but as Simpson suggests, they are a good start to hold companies accountable to. We might insist, for example, that engagement algorithms must provide results that are objective and evidence-based in addition to seeking usage-maximization. Following Fricker, epistemic values understood this way are actually *moral* values, since not engaging them harms us in our capacities as knowers. Requiring companies to be legally bound by these other values would make them liable, open to lawsuits, if they step out of line. It should be noted, though, that such lawsuits, even fines in the millions of dollars, are often just part of operating costs for sufficiently large companies; holding them legally accountable may be more effective in shaping how the public thinks of the role of social media companies in society than it is as a financial deterrent.

On a more grassroots level, public policy circles have popularized the term "countering violent extremism", or CVE for short (see for example Lee 2019). Among the strategies employed in CVE is "counter-messaging" to try to dissuade those caught in radicalization pipelines from proceeding further. The main criticism against CVE is a criticism I am sympathetic to, which is as follows: how can it deliver a message to its audience? And, even if they are able to successfully deliver a message to people radicalized online, why would they listen?

The Bipartisan Policy Center's 2012 Homeland Security Project report provides us with an effective summary of government strategies to combat radicalization:

> "Approaches aimed at restricting freedom of speech and removing content from the Internet are not only the least desirable strategies, they are also the least effective. Instead, government should play a more energetic role in reducing the demand for radicalization and violent extremist messages—for example, by encouraging civic challenges to extremist narratives and by promoting awareness and education of young people.
>
> In the short term, the most promising way to deal with the presence of violent extremists and their propaganda on the Internet is to exploit, subject to lawful guidelines and appropriate review and safeguards, their online communications to gain intelligence and gather evidence in the most comprehensive and systematic fashion possible." (1)

In other words, the BPC recommends a focus on reducing the *demand*, rather than reducing the *supply.* Recall from the introduction that part of the nature of radicalization is not to increase the supply of misinformation, but the demand for it. This still favours a "marketplace of ideas" approach, which is subject to the same limitations as the "epistemic virtue" argument is, but we should be sympathetic to the argument that reducing the supply is ineffective and not desirable. Any attempt to remove content from the Internet is likely to be a serious breach of freedom of speech. The BPC's document goes on to outline just what the aims of a government should be in combating online

radicalization:

> "• Government, in partnership with community groups, needs to continue to expand programs and initiatives that create awareness and spread information about online radicalization among educators, parents, and communities.
> • Government should serve as an enabler, bringing together the private sector, foundations, philanthropists, and community groups to build capacity and to help potentially credible messengers—such as mainstream groups, victims of terrorism, and other stakeholders—to become more effective at conveying their messages. The forthcoming Internet strategy should spell out what the government will do and how success will be measured.
> • The government's Internet strategy also needs to make clear what part of government will coordinate capacity building, engagement, and outreach efforts as well as what resources will be made available to support this task.
> • The government should encourage school authorities to review and update their curricula on media literacy, consider violent extremism as part of their instruction on child-safety issues, and develop relevant training resources for teachers." (1)

Our takeaway

What, finally, to make of all this? I think we have good reason to be conservatively in favour of "epistemic vigilance", simply because it does not cost us much to accept it, with the caveat that this is not much more than a first line of defense to protect already non-radicalized agents from becoming radicalized. Again, as with CVE,

the difficulty for such solutions is always finding some way to deliver them to radicalized people. To reiterate, the suggestion of "epistemic vigilance" places the onus of good values and good, responsible behaviour on the user. This is a bitter pill to swallow; however, it might be a better solution to teach individual citizens how to guard themselves online than to instruct corporations to bind themselves by such-and-such values in an ever-evolving market and society. In my (2021b) manuscript I argue that "following Néron and Norman (2008), perhaps we don't really want corporations with free access to our data to try to be paragons of responsibility, virtue and "wokeness". It seems right that we should maintain a healthy dose of skepticism towards these monoliths, especially when they are eager to earn our trust." (8)

Beyond vigilance, I think we have good reason to be in favour of government intervention. Despite de Melo-Martín and Intemann's concerns that government intervention can be heavy-handed, stifling the wrong kind of discourse, and also runs the risk of stirring up claims of censorship, I think it provides the surest path to combating online radicalization. Whether this intervention is limited, focusing mainly on public mediation, or the "light touch" of changing personalization defaults as Simpson suggests, or heavier, focusing on legislation as Fritts and Cabrera suggest, or even as far as a series of antitrust suits, I believe this provides the most promising avenue of change in the ongoing fight against radicalization. Legally binding companies to replace or balance the value of usage-maximization with other values in the operation of engagement algorithms may not make much of a dent in their bottom line, but it would, at the very least, go some distance in changing public perception of their operation.

Finally, I think we have good reason to be against allowing corporations to police

themselves, and to spearhead the implementation of solutions to online radicalization. Following reports such as The Wall Street Journal's "The Facebook Files", tech monoliths have already shown what they think of data ethics, and it's not reasonable to assume that anything would change if they were given more responsibility in this regard.

Concluding remarks

Contributions

Throughout the thesis, I have made several contributions to the philosophical literature of online radicalization. First, I have shown that online radicalization, as accelerated by social media algorithms, problematizes a popular picture of radicalization. The popular picture we have looked at, exemplified by such documents as those by the Bipartisan Policy Center and the Canada Centre for Community Engagement and Prevention of Violence, implies that a radicalized agent embody traits that I hope to have shown are not required – for example, that they be social outcasts, that they struggle with mental health issues, that they are (or become) strongly disposed to commit physical violence, and that their radicalization occurs on a timeline with discernible points.

Second, I have shown that Fricker's characterization of epistemic injustice – especially the characterization of epistemic harm that features within – can illustrate the harm done to an agent caught in engagement algorithm-accelerated online radicalization.

Finally, I have shown that engagement algorithm-accelerated online radicalization is plausibly distinct from offline radicalization because of the accelerated epistemic harm it can cause to an agent.

Other avenues

There is much more to be said on the topic of online radicalization that has not been covered in these pages. Public policy circles, drawing particularly on behavioural science, have said a great deal in recent years about online radicalization, though their characterizations are still heavily weighted on the concept of the "lone wolf" mass shooter that I have problematized throughout.

An entire “meme theory” dissertation could be written on the ever-changing nature of memes and their role in online radicalization, especially on YouTube and TikTok. The radicalizing power of memes lies largely in their ability to code information without the use of language, such that they can convey objectionable content but never be outright challenged, since to speak out against them is to face an endless moving of goalposts (“that’s not what it meant”, “you just don’t get it”, etc.).

Social media and Internet use quite plausibly has some sort of effect on our brain chemistry. It follows that our altered brains might act, think and form beliefs differently. Research in this field might give neurological credence to my adaptation of Fricker’s argument that social media harms us in our capacities as knowers. For example, Macit, Macit & Güngör (2018) and Burhan & Moradzadeh (2020) found that use of social media created a dopamine feedback loop in the user, eventually developing a dependence on social media use. Schmack et al. (2015) found that “greater dopamine availability was associated with a stronger propensity towards unfounded beliefs”. Nour et al. (2018) found that dopamine played a role in updating an agent’s belief model of the world, and believed their findings had “implications for understanding the pathophysiology of psychotic disorders where dopamine function is disrupted.” A great deal more research would be needed, but I believe it’s plausible that the social media dopamine feedback loop has an adverse effect on belief formation: to wit, that users can become chemically addicted to reading controversial opinions online, and the feelings they receive from engaging with such opinions.

Limitations

This work has focused only on major Western social media platforms that the author has familiarity with – in particular, Facebook, Instagram, YouTube, Twitter, and to some limited degree, TikTok. It says very little about social media platforms that are popular in other countries and with other age demographics – for example, WhatsApp, KakaoTalk, VK and WeChat. It is unknown how these platforms contribute to the online radicalization of some of their users; cultural norms and legislated freedom of speech in the relevant countries where they are popular, as well as ability to access and interact with various features of those apps, would likely all play some kind of role in modifying the degree to which these platforms manipulate their users.

References

Alfano, Mark; Carter, Joseph Adam & Cheong, Marc (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association* (3):298-322.

Alfano, Mark; Fard, Amir Ebrahimi; Carter, Joseph Adam; Clutton, Peter & Klein, Colin (2020). Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*:1-24.

Alfano, S. (2009, February 11). *The truth of truthiness.* CBS News. http://www.cbsnews.com/stories/2006/12/12/opinion/meyer/main2250923.shtml

Basch, C.H., Meleo-Erwin, Z., Fera, J., Jaime, C. & Basch, C.E. (2021) A global pandemic in the time of viral memes: COVID-19 vaccine misinformation and disinformation on TikTok, Human Vaccines & Immunotherapeutics, 17:8, 2373-2377, DOI: 10.1080/21645515.2021.1894896

Bridgman, A., Merkley, E., Zhilin, O., Loewen, P. J., Owen, T., & Ruths, D. (2021). Infodemic Pathways: Evaluating the Role That Traditional and Social Media Play in Cross-National Information Transfer. In Frontiers in Political Science (Vol. 3). Frontiers Media SA. https://doi.org/10.3389/fpos.2021.648646

Burhan, R. and Moradzadeh, J., Neurotransmitter Dopamine (DA) and its Role in the Development of Social Media Addiction. *Journal of Neurology & Neurophysiology*, 2020, 11(7), 01-02.

Burke, Liam (2021a). Engagement algorithms and epistemic vulnerability. Unpublished manuscript submitted as final paper, Carleton University.

Burke, Liam (2021b). Value-neutrality, value-blindness and plausible corporate

deniability. Unpublished manuscript submitted as final paper, Carleton University.

Canada Centre for Community Engagement and Prevention of Violence. *National strategy on countering radicalization to violence.* Public Safety Canada. https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/ntnl-strtg-cntrng-rdclztn-vlnc-en.pdf

Charles, A. (2013, May). *How liking YouTube videos works.* Chron Small Business. https://smallbusiness.chron.com/liking-youtube-videos-works-67700.html

Coady, C.A.J. (1995). *Testimony: A Philosophical Study*. Clarendon.

Côté-Bouchard, C. (2020, September). "Is the internet safe? Personalization and the threat of epistemological skepticism." Poster session presented at the fall colloquium of Carleton University, Ottawa.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh, Pa: University of Pittsburgh Press. Retrieved March 1, 2021, from http://www.jstor.org/stable/j.ctt6wrc78

European College of Neuropsychopharmacology(2018, October 10). *Researchers Show Change in Beliefs Associated with Dopamine in Brain.* Neuroscience News. https://neurosciencenews.com/belief-dopamine-9990/

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. New York, Ny: Oxford University Press.

Fritts, Megan & Cabrera, Frank (forthcoming). Fake news and epistemic vice: combating a uniquely noxious market. *Journal of the American Philosophical Association*.

Gow, P. and Rookwood, J. (2008) Doing it for the team - examining the causes of

contemporary English football hooliganism. *Journal of Qualitative Research in Sports Studie*s 2 (1):71-82.

Homeland Security Project. *Countering online radicalization in America: Executive summary.* Bipartisan Policy Center. https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2019/03/5086_BPC-_Online-Radicalization-Report-Executive-Summary-v4_web.pdf

Hussain, W., & Moriarty, J. (2016). Accountable to Whom? Rethinking the Role of Corporations in Political CSR. *Journal of Business Ethics* (Vol. 149, Issue 3, pp. 519–534). Springer Science and Business Media LLC. https://doi.org/10.1007/s10551-016-3027-8

Ismail, S. (2018, May 14). *Why algorithms are the future of business success.* The Growth Institute. https://blog.growthinstitute.com/exo/algorithms

Lee, B. (2019). Countering Violent Extremism Online: The Experiences of Informal Counter Messaging Actors. In Policy & Internet (Vol. 12, Issue 1, pp. 66–87). Wiley. https://doi.org/10.1002/poi3.210

Letzing, J. and Berkley, A. (2021, July 13). *Is the internet really more effective at radicalizing people than older media?* World Economic Forum. https://www.weforum.org/agenda/2021/07/is-the-internet-really-more-effective-at-radicalizing-people-than-older-media/

Martin, J. (2020, November 14). *Reddit co-founder calls out social media for spreading conspiracies: "we're gonna have to deradicalize a lot of people"*. Newsweek. https://www.newsweek.com/reddit-co-founder-calls-out-social-media-spreading-

conspiracies-were-gonna-have-deradicalize-1547413?piano_t=1

Macit, H. B., Macit, G., & Güngör, O. (2018). SOSYAL MEDYA BAĞIMLILIĞI VE DOPAMİN ODAKLI GERİBİLDİRİM ÜZERİNE BİR ARAŞTIRMA. *Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 882–897. https://doi.org/10.30798/makuiibf.435845

Masnick, M. (2003, February 11). *Companies try infotising sites.* Techdirt. https://www.techdirt.com/2003/02/11/companies-try-infotising-sites/

McCluskey, M. (2021, September 15). From Instagram's toll on teens to unmoderated 'elite' users, here's a break down of the Wall Street Journal's Facebook revelations. *Time.* https://time.com/6097704/facebook-instagram-wall-street-journal/

Melo-Martín, I., & Intemann, K. (2018). *The fight against doubt : how to bridge the gap between scientists and the public*. Oxford University Press.

Miller, Boaz & Record, Isaac (2013). Justified belief in a digital age: on the epistemic implications of secret internet technologies. *Episteme* 10 (2):117 - 134.

Néron, P.-Y., & Norman, W. (2008). Citizenship, Inc.: Do We Really Want Businesses to Be Good Corporate Citizens? *Business Ethics Quarterly*, 18(1), 1–26. http://www.jstor.org/stable/27673212

Neumann, P. R. (2013). Options and Strategies for Countering Online Radicalization in the United States. In Studies in Conflict & Terrorism (Vol. 36, Issue 6, pp. 431–459). Informa UK Limited. https://doi.org/10.1080/1057610x.2013.784568

Newberry, C. (2022, February 28). *How the Facebook algorithm works in 2022 and how to make it work for you.* Hootsuite. https://blog.hootsuite.com/facebook-

algorithm/

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H. B., Coello, C., Wall, M. B., Dolan, R. J., & Howes, O. D. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. Proceedings of the National Academy of Sciences of the United States of America, 115(43), E10167–E10176. https://doi.org/10.1073/pnas.1809298115

Nguyen, C. Thi (2020). Echo chambers and epistemic bubbles. *Episteme* 17 (2):141-161.

Perreaux, L. and Andrew-Gee, E. (2017, January 30). Quebec City mosque attack suspect known as online troll inspired by French far-right. *The Globe and Mail.* https://www.theglobeandmail.com/news/national/quebec-city-mosque-attack-suspect-known-for-right-wing-online-posts/article33833044/

Satz, D. (2010). *Why Some Things Should Not Be for Sale: The Moral Limits of Markets.* New York: Oxford University Press.

Schmack, K., Rössler, H., Sekutowicz, M., Brandl, E. J., Müller, D. J., Petrovic, P., & Sterzer, P. (2015). Linking unfounded beliefs to genetic dopamine availability. Frontiers in Human Neuroscience, 9, 521. https://doi.org/10.3389/fnhum.2015.00521

Simpson, Thomas W. (2012). Evaluating Google as an epistemic tool. *Metaphilosophy* 43 (4):426-445.

Statt, N. (2020, May 26). *Facebook reportedly ignored its own research showing algorithms divided users.* The Verge. https://www.theverge.com/2020/5/26/21270659/facebook-division-news-feed-algorithms

Sullivan, Emily, Sondag, Max, Rutter, Ignaz, Meulemans, Wouter, Cunningham, Scott, Speckmann, Bettina & Alfano, Mark (2020). Vulnerability in social epistemic networks. *International Journal of Philosophical Studies* 28 (5):1-23.

The Wall Street Journal. (n.d.). *The Facebook files.* https://www.wsj.com/articles/the-facebook-files-11631713039

Youth.gov. (n.d.). *Online safety for youth: working to counter online radicalization to violence in the united states.* https://youth.gov/feature-article/online-safety-youth-working-counter-online-radicalization-violence-united-states