6

Epistemic Paternalism and Social Media

In the previous chapters, we explored various ways in which algorithms can undermine our freedom or threaten our autonomy, paying particular attention to the context of the prison sentencing and employee management (specifically, teachers and Uber drivers). In this chapter, we turn to a different context – which we briefly touched on in the previous chapter – that of our current media environment, which is in large part shaped by algorithms. We discuss some distinctively epistemic problems that algorithms pose in that context and some paternalistic solutions they call for. Our paternalistic proposal to these problems is compatible with respect to freedom and autonomy; in fact, our freedom and autonomy demand them.

Let us begin with some reflections on our current media environment. In 1995, MIT media lab founder Nicholas Negroponte foresaw a phenomenon that we are now all familiar with: the replacement of traditional newspapers with virtual newspapers, custom-fitted to each reader's particular taste. In his speculations, he called the virtual newspaper "the Daily Me." Cass Sunstein elaborates on the idea of the Daily Me:

> Maybe your views are left of center, and you want to read stories fitting with what you think about climate change, equality, immigration, and the rights of labor unions. Or maybe you lean to the right, and you want to see conservative perspectives on those issues, or maybe on just one or two, and on how to cut taxes and regulation, or reduce immigration. Perhaps what matters most to you are your religious convictions, and you want to read and see material with a religious slant (your own). Perhaps you want to speak to and hear from your friends, who mostly think as you do, you might hope that all of you will share the same material. What matters is that with the Daily Me, everyone could enjoy an *architecture of control*. Each of us would be fully in charge of what we see and hear.[1]

As Negroponte anticipated, custom-fitted virtual news has become widespread and popular. This has been facilitated by the advent of "new media" – highly interactive digital technology for creating, sharing, and consuming information. New media is

[1] Sunstein, *#Republic: Divided Democracy in the Age of Social Media*, 1 (emphasis added).

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

now pervasive, with more Americans getting their news from social media (the predominant form of new media) than traditional print newspapers.[2]

In 1995, the Daily Me might have sounded like a strict improvement on traditional news. However, we now know that the architecture of control that it affords us has serious drawbacks. Consider an episode that briefly caught the nation's attention in the summer of 2019. About a month after the Mueller Report – the *Investigation into Russian Interference in the 2016 Presidential Election*[3] – was released, Justin Amash (who was a Republican member of the U.S. House of Representatives) gave a town hall to explain why he thought the report was grounds for impeaching the president. At the town hall, NBC interviewed a Michigan resident who stated that she was "surprised to hear there was anything negative in the Mueller Report *at all* about [the President]"[4] (Golshan 2019; emphasis added). At the time, it was hard to see how anyone could think this. Yet she thought the report had exonerated the president.

When the resident learned that the report contained negative revelations about the president, it was through serendipity. Amash was the only Republican representative calling for impeachment following the release of the report, and she happened to live in his district. Had it not been for this, she likely would have continued to believe that report had exonerated the president.

The Michigan resident is not a special case. Many people continue to believe that the Mueller Report explicitly concludes that the president and members of his campaign did nothing wrong. Moreover, the phenomenon of people being misinformed in similar ways is common. Along with the architecture of control afforded by customized news comes the danger of encapsulating ourselves in "epistemic bubbles," epistemic structures that leave relevant sources of information out and walling ourselves off in "echo chambers," epistemic structures that leave relevant sources of information out, and actively discredit those sources.[5] This is, in part, due to the automated way in which news feeds and other information-delivery systems (such as search results) are generated. Eli Pariser explains:

> The new generation of Internet filters look at the things you seem to like [. . .] and tries to extrapolate. They are prediction engines, constantly creating and refining a theory of who you are and what you'll do and want next. Together, these engines create a unique universe of information for each of us and [. . .] alters the way we encounter ideas and information.[6]

[2] Shearer, "Social Media Outpaces Print Newspapers in the U.S. as a News Source."

[3] U.S. Department of Justice, "Report on the Investigation into Russian Interference in the 2016 Presidential Election, Volume I ('Mueller Report')"; U.S. Department of Justice, "Report on the Investigation into Russian Interference in the 2016 Presidential Election, Volume II ('Mueller Report')."

[4] Caldwell and Moe, "Republican Justin Amash Stands by Position to Start Impeachment Proceedings" (emphasis added).

[5] Nguyen, "Echo Chambers and Epistemic Bubbles."

[6] Pariser, *The Filter Bubble*, 9.

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

This is why cases like the Michigan resident are far from isolated: Many of us are in epistemic bubbles or echo chambers built through the cooperation of internet filters and ourselves.

Consider the user interface of YouTube, the world's largest social media platform.[7] The site's homepage opens to a menu loaded up with algorithmically generated options, curated to match each user's tastes. Immediately after watching any video on the homepage, users are met with algorithmically generated suggestions based on what they have just watched. The suggestions that appear on the home page and at the conclusion of videos are, of course, chosen to keep users on the site. So users who are inclined toward, say, left-of-center politics are more likely to receive suggestions for videos supporting that worldview. Given that two-thirds of Americans get news through social media – with one in five getting news from YouTube[8] – it is no wonder we can consume a lot of news but be mis- or under-informed about current affairs of monumental importance.

New media's design and popularity also facilitate the mass spread of misinformation. This is not only unfortunate but dangerous. Consider the recent surge in "vaccine hesitancy" – the reluctance or refusal to vaccinate – a phenomenon that the World Health Organization now considers one of the top ten threats to global health.[9] The surge of vaccine hesitancy seems to be inseparable from the success of new media, with Facebook, Twitter, and YouTube playing a large role in spreading misinformation about vaccines.[10] Now, decades after it had been declared "eliminated," measles has returned to the United States.[11] And in the recent COVID-19 pandemic, one early concern was that the low rates of seasonal flu vaccine would increase strain on health-care systems both by requiring more testing to differentiate COVID-19 and seasonal flu cases and by increasing strain on treatment resources.

This raises questions about the responsibility new media developers have to manage the architecture of control that its users currently enjoy. It also raises questions about the latitude that social media developers have in making alterations to their sites. On the one hand, it seems reasonable to think that developers should lead their users to consider a more diverse array of points of view, even if that is not in line with users' immediate wishes. On the other, there seems to be something objectionably paternalistic about this: Users should be able to (at least in some sense) decide their information diets for themselves.

We will argue that there is plenty of room for epistemic paternalism online. Moreover, because the internet information environment is epistemically noxious, such epistemically paternalistic policies should be a persistent part of the internet information environment. This chapter proceeds as follows. First, we discuss an

7 Kaiser and Rauchfleisch, "Unite the Right?"
8 Shearer and Matsa, "News Use across Social Media Platforms 2018."
9 World Health Organization, "Ten Health Issues WHO Will Tackle This Year."
10 Hussain et al., "The Anti-Vaccination Movement."
11 Joy, "What's Causing the 2019 Current Measles Outbreak?"

intervention that Facebook has run in hopes of demoting the spread of fake news on the site. We explain why the intervention is paternalistic and then, using the framework of this book, defend the intervention. We argue that while Facebook's intervention is defensible, it is limited. It is an intervention that may pop some epistemic bubbles but will likely be powerless against echo chambers. We then discuss heavier-handed interventions that might be effective enough to dismantle some echo chambers, and we argue that at least some heavier-handed epistemically paternalistic interventions are permissible.

6.1 DEMOTING FAKE NEWS

In April 2019, Facebook announced that it would use a new metric, Click-Gap, to determine where to rank posts in its users' News Feeds.[12] Click-Gap measures the gap between a website's traffic from Facebook and its traffic from the internet at large, and it demotes sites with large gaps. According to Facebook, the idea is that "a disproportionate number of outbound Facebook clicks [. . .] can be a sign that the domain is succeeding on News Feed in a way that doesn't reflect the authority they've built outside it."[13] Click-Gap attempts to identify and demote low-quality content, such as fake news, in News Feed to prevent it from going viral on the website.[14]

We will argue that Click-Gap is an instance of epistemic paternalism and that it is morally permissible. We begin by explaining why we take Click-Gap to be an instance of epistemic paternalism. We then argue that it is permissible, despite its being paternalistic. It may seem strange that in a book about the threat that algorithms can pose to autonomy that we would take this line. As we will soon argue, however, the epistemically noxious online information environment creates a need for action. And, whether such action is paternalistic or executed algorithmically does not matter *as such*. Instead, what matters is whether these actions are consistent with respect for persons, properly understood. We will demonstrate that Click-Gap – and a host of other potential interventions – occupies this exact space. Certain paternalistic interventions, like Click-Gap, do not undermine users' autonomy; they, in fact, support it.

Let's begin by discussing our understanding of paternalism, as it deviates from the standard philosophical understanding of the concept. A standard conception of paternalism is as follows:

> **Paternalism:** *P* (for "paternalist") acts paternalistically toward *S* (for "subject") by *φ*'ing (where "*φ*" denotes an action) *iff* the following conditions are met:

[12] "News Feed is a personalized, ever-changing collection of photos, videos, links, and updates from the friends, family, businesses, and news sources you've connected to on Facebook." Facebook, "News Feed."

[13] Rosen and Lyons, "Remove, Reduce, Inform."

[14] Rosen and Lyons.

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

Interference: φ'ing interferes with the liberty or autonomy of *S*.
Non-Consent: *P* does so without the consent of *S*.
Improvement: *P* does so just because *φ*'ing will improve the welfare of *S* (where this includes preventing *S*'s welfare from diminishing), or in some way promote the interests, values, or good of *S*.[15]

As we will soon argue, the interference condition is not met in the case of Click-Gap. Yet, as we have already noted, we take it that the intervention is an instance of paternalism. This is because we reject the interference condition.

Let us next explain why Click-Gap does not meet the interference condition. We take it that agents are *autonomous* when they enjoy procedural independence (i.e., when they are competent and their beliefs, preferences, desires, and values are authentic) and substantive independence (i.e., when their procedural independence is supported by their social and relational circumstances). We understand agents as *free* when they are undominated, autonomous, and sufficiently effective as agents. To show that Click-Gap does not meet the interference condition, then, we will show that the case does not, by the lights of these definitions, undermine freedom, autonomy, or quality of agency.

First consider autonomy. If Click-Gap is successful, it will influence the attitudes of Facebook users – some, for example, will adopt different views about vaccine safety than they otherwise would have. This influence is not enough, though, to undermine users' autonomy. The relevant question for the purposes of assessing their autonomy is whether the attitudes the users adopt will be authentic – which is to say that those attitudes are ones that they would endorse upon critical reflection as consistent with their beliefs and values over time – not simply whether they were influenced in some way or other. And Click-Gap influences users by shielding them from content that may seem more credible than it actually is because it rose to the top of News Feed by gaming the News Feed algorithm. Presumably, the attitudes formed in the absence of such manipulation and misinformation are the sort agents can authentically endorse. Click-Gap, then, will *prevent* users from forming inauthentic attitudes. This is owed to the fact that people (at least typically) will desire that their beliefs be justified and accurate.

Fair enough, one might say, but what about users who do not care about the truth and want to be anti-vaxxers, come what may. Wouldn't these – perhaps fictional – users have their autonomy undercut by Click-Gap if it changed their view? For all we've said, it is possible that these users would have their autonomy undercut *if Click-Gap changed their views*. But we are skeptical that Click-Gap could have an effect on such users. Click-Gap is not responding to the content of claims made by anti-vaxxers, it is simply demoting their posts. Moreover, de-emphasizing patently bad information (or misleading information) that happens to confirm antecedent views does not undermine those views per se. Rather, it mitigates unjustified

[15] Dworkin, "Paternalism."

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

confirmation. In the end, a committed anti-vaxxer isn't likely to change their mind as a result of Click-Gap.[16] Its touch is far too light for that.

But what about a less narrowly defined group, such as committed skeptics of the medical establishment? Might they have a complaint? Again, either such skeptics are dogmatic skeptics, who will be skeptical come what may, in which case they are not likely to have their minds changed, or they are not, in which case Click-Gap will be – if anything – aiding them in their pursuit of the truth by filtering out low-quality information.

Let's now turn to freedom. Some users may not *like* what Click-Gap does, but it does not dominate them, rob them of resources to effectuate their desires, or diminish their capacity to act as agents. Under the policy, users can still post what they were able to post before, follow whomever they were able to follow before, and so on. Now, Click-Gap *does* interfere with the freedom of purveyors of low-quality content of Facebook, but that is a separate matter (one that we will deal with later). We are not arguing that Click-Gap impedes no one's freedom; we are arguing that the freedom it *does* impede is not socially valuable.

This raises a complication: Since purveyors are users, one might argue that the interference condition *does* apply in the case of Click-Gap. It is true that the interference condition is met in this special case, but this is beside the point. In the special case where the interference condition is met, the improvement condition (the condition that the intervention is taken because it will improve the welfare of the person whose freedom or autonomy is affected) is *not* met. The intervention is not taken to shield those sharing low-quality content from their own content. Rather, it is to shield potential recipients from that information.

So Click-Gap does not meet the interference and improvement conditions simultaneously. Hence, it would not be an instance of paternalism on the standard account.

But the standard account of paternalism has important limitations. The view of paternalism we wish to advance here addresses those limitations. On that conception, Click-Gap *is* an instance of paternalism. The reason we do not adopt the standard definition of paternalism is because the improvement condition itself is flawed. To see why, consider Smoke Alarm:

> **Smoke Alarm.**[17] Molly is worried about the safety of her friend, Ray. Molly knows that there is no smoke alarm in Ray's apartment and that he tends to get distracted while cooking. Molly thinks that if she were to suggest that Ray get one, he would agree. But – knowing Ray – she does not think he would actually get one. She thinks

[16] This is not to say that the architecture of social media sites cannot influence users in important ways. We take it that the "technological seduction" that sites like YouTube exhibit *can* encroach on autonomy by, for example, seeding and nurturing convictions that either cannot be endorsed upon reflection *or* have been seeded and nurtured through methods that agents are alienated from. See Alfano, Carter, and Cheong, "Technological Seduction and Self-Radicalization."

[17] This case is inspired by an example from Ryan, "Paternalism: An Analysis."

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

that if she were to offer to buy him one, he would refuse. She buys him one anyway, thinking that he will accept and install the already bought alarm.

Molly's gifting a smoke alarm does not meet the interference condition. Ray adopts no attitudes that he is alienated from and he is left free to do what he wills. Yet – intuitively – Molly acts paternalistically toward him, at least to the extent that she acts to contravene his implicit choice not to install a smoke alarm in his apartment.

Our primary reason for rejecting the standard definition is its susceptibility to counterexamples like Smoke Alarm. But there are other issues. As Shane Ryan convincingly argues, there are also issues with the non-consent and improvement conditions.[18]

Contra the non-consent condition (the condition that *P φ's* without the consent of *S*): we seem to be able to act paternalistically, even when the paternalism is welcomed. Ryan drives this point with the example of a Victorian wife who has internalized the sexist norms of her culture and wills that her husband makes her important decisions for her.[19] The Victorian wife's husband's handling of her issues is paternalistic, even though she consents to it.

Contra the improvement condition (the condition that *P φ's* just because it will improve the welfare of *S*), we can act paternalistically when we fail to improve anyone's welfare.[20] Suppose the new alarm lulls Ray into a false sense of security, which results in even more careless behavior and a cooking fire. This is not enough to show that Molly's gesture was not paternalistic. All that seems to be required is that she *thought* that buying the smoke alarm would make him better off. Whether it does is beside the point.

For these reasons, we adopt the following account of paternalism from Ryan (2016):

Paternalism*: *P* acts paternalistically toward *S* by *φ*'ing *iff* the following conditions are met:

Insensitivity: *P* does so irrespective of what *P* believes the wishes of *S* might be.
Expected improvement: *P* does so just because *P* judges that *φ*'ing might or will advance *S*'s ends (*S*'s welfare, interests, values or good).[21]

By this definition, Click-Gap could qualify as a paternalistic intervention, so long as it is motivated by Facebook's judgment that improving people's epistemic lot improves their welfare.

Now, the connection between one's epistemic lot and their welfare is contested and complicated. It is certainly true that in many instances having knowledge – say, about where the lions, tigers, and Coronaviruses are, and thus how to avoid them – is often conducive to welfare. But it is not clear that knowledge is *always* conducive to welfare. Sometimes knowledge is irrelevant to our ends: Consider Ernest Sosa's

[18] Ryan.
[19] Ryan.
[20] Ryan.
[21] Ryan.

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

example of knowing how far two randomly selected grains of sand in the Sahara are from one another.[22] Likewise, sometimes knowledge can be detrimental to our ends: Consider Thomas Kelly's example of learning how a movie ends before you see it,[23] or Bernard Williams's case of a father whose son was lost at sea but for his own sanity believes – however improbably – that his son is alive somewhere.[24]

For these reasons, we will couch the rest of our discussion in terms of *epistemic* paternalism:

> **Epistemic Paternalism:** *P* acts epistemically paternalistically toward *S* by *φ'ing iff* the following conditions are met:
>
> *Insensitivity*: *P* does so irrespective of what *P* believes the wishes of *S* might be.
> *Expected epistemic improvement*: *P* does so just because *P* judges that *φ*'ing might or will make *S* epistemically better off.

Following Kristoffer Ahlstrom-Vij we will understand agents as epistemically better off when they undergo epistemic Pareto improvements with respect to a question that is of interest to them (where *epistemic Pareto improvements* are improvements along at least one epistemic dimension of evaluation without deterioration with respect to any other epistemic dimension of evaluation).[25]

Despite failing to meet the traditional conception of paternalism, Click-Gap is an instance of epistemic paternalism. Adam Mosseri, Vice President of Facebook's News Feed, has stated that Click-Gap is part of an effort to make users better informed.[26] In other words, the policy is in place so as to make users epistemically better off (i.e., expected epistemic improvement is met). Mosseri takes it that Facebook has an obligation to stage interventions like Click-Gap in order to fight the spread of misinformation via Facebook products. In explaining this obligation, Mosseri states that "all of us – tech companies, media companies, newsrooms, teachers – have a responsibility to do our part in addressing it."[27] The comparison to teachers is apt. Teachers often must be epistemically paternalistic toward their students. That is, they often must deliver information irrespective of the wishes of their students, just because it will epistemically benefit their students. Like teachers, Facebook won't tailor its delivery of information to the exact wants of its constituency. That is, it won't abandon Click-Gap in the face of pushback or disable it for users who do not want it. So insensitivity is met. This is, in part, due to the kind of pressure Facebook is reacting to when it takes part in interventions like Click-Gap, such as pressure from the public at large and governments[28] to fight the spread of fake news.

[22] Sosa, "For the Love of Truth?"
[23] Kelly, "Epistemic Rationality as Instrumental Rationality: A Critique."
[24] Williams, "Deciding to Believe."
[25] Ahlstrom-Vij, *Epistemic Paternalism*.
[26] Mosseri, "Working to Stop Misinformation and False News."
[27] Mosseri.
[28] Germany is proposing a law to fine Facebook for advertisements containing fake news. See Olson, "Germany Wants Facebook to Pay for Fake News."

Now that we have articulated why Click-Gap is epistemically paternalistic, let's now turn to the moral question: Is it permissible?

We will address this question by exploring it from two vantage points: that of Facebook users and purveyors. Let's start by looking at the intervention from the perspective of Facebook users. What claims might users have against the policy? We have already argued that the policy is not a threat to autonomy or freedom. Further, it is not plausible that the intervention will harm users. Given this – and the fact that the intervention is driven by noble aims – it is hard to see how users could reasonably reject Click-Gap.

What about the purveyors?

They might claim that, unlike users, this policy does undermine their autonomy or freedom. But it is difficult to see how much (if any) weight can be given to this claim. Start with autonomy. Purveyors are not fed any attitudes from which they could be alienated or which undermine their epistemic competency, so any claims from procedural independence will lack teeth. Claims from substantive independence will also miss the mark. Limiting persons' ability to expose others to misinformation does nothing to undermine means of social and relational support or create impediments to one's ability to exercise de facto control over their life. Hence, this kind of epistemic paternalism undermines neither psychological nor personal autonomy.

Complaints from freedom fail in similar fashion. Click-Gap does introduce constraints on effectuating the desires of purveyors. But more is needed to show that this makes Click-Gap *wrong*. The claim that it is wrong to place any constraint on freedom is clearly false, and one our Chapter 5 account rejects. Our account says that morally considerable freedom is quality of agency, and one's quality of agency is not diminished by limitations on the ability to disseminate misinformation. That's because exercising autonomy requires the ability to advance one's interests and abide fair terms of social cooperation. Sullying the epistemic environment may advance one's interests, but it is not consonant with fair terms of social cooperation. Further, purveyors still are left otherwise free to promote their ideas on the site. One way of doing this – posting content that does well on Facebook and not elsewhere – has been made less effective, but they are left free to promote their ideas by other means.

Finally, purveyors might say they have an interest impeded by the intervention. Given that their content is still allowed on Facebook and that Click-Gap leaves open many avenues to promoting content, the interest is only somewhat impeded. Moreover, it is not a particularly weighty interest. And even though the interest is impeded, there is an open question of whether such impediments are justified. We think they are, once we consider the reasons that speak in favor of the intervention. We turn to those next.

It is reasonable to think that Click-Gap will prevent significant harms to individuals. Policies like this have been proven to be quite effective. Consider Facebook's

2016 update to the Facebook Audience Network policy, which banned fake news from ads.[29] As a result of this ban, fake news shared among users fell by about 75 percent.[30] Since fake news is a driver of harmful movements, such as vaccine hesitancy, there is a strong consideration in favor of the policy. After all, those who wind up sick because of vaccine hesitancy are significantly harmed.

Harms are not the only things to consider. Policies like Click-Gap also support user autonomy. Consider the following:

> A growing body of evidence demonstrates that consumers struggle to evaluate the credibility and accuracy of online content. Experimental studies find that exposure to online information that is critical of vaccination leads to stronger anti-vaccine beliefs, since individuals do not take into account the credibility of the content [. . .]. Survey evidence [. . .] shows that only half of low-income parents of children with special healthcare needs felt "comfortable determining the quality of health websites" [. . .]. Since only 12% of US adults are proficient in health literacy with 36% at basic or below basic levels [. . .], Fu et al. (2016) state that [. . .] "low-quality antivaccine web pages [. . .] promote compelling but unsubstantiated messages [opposing vaccination]."[31]

Interventions like Click-Gap are an important element of respecting users' autonomy. The intervention, if successful, will protect users from internalizing attitudes that would be inauthentically held.

Click-Gap is thus a policy that all parties effected could reasonably endorse. Further, Facebook *must* engage policies like Click-Gap: Users who adopt unwarranted beliefs because of fake news and individuals who contract illnesses because of vaccine hesitancy have a very strong claim against Facebook's taking a laissez-faire approach to combating fake news on its site.

Interventions like Click-Gap, then, are not only permissible; they should be common. Such interventions, however, are limited. Click-Gap might be able to pop some users' epistemic bubbles, but they are unlikely to dismantle sturdier structures such as echo chambers. So there is good reason to look into what we may do to chip away at these structures, a topic to which we now turn.

6.2 DISMANTLING ECHO CHAMBERS

In fall 2017, Reddit – a social media site consisting of message boards ("subreddits") based on interests (e.g., science, world news, gaming) – banned r/incels,[32] a message board for "incels," people who have trouble finding romantic partners.[33] The group

[29] Wingfield, Isaac, and Benner, "Google and Facebook Take Aim at Fake News Sites."
[30] Chiou and Tucker, "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement."
[31] Chiou and Tucker.
[32] The URL of a subreddit begins with "reddit.com/r/." For this reason, many subreddits, such as the subreddit for world news, are referred to as "r/worldnews."
[33] Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."

was banned for hosting "content that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people."[34] We take it that the ban was justified; banning a group is an intrusive step, but such intrusions can be justified when the stakes are high, such as when physical harm is threatened. Here, we would like to ask what Reddit may have done before the stakes were so high and what Reddit may have done before things got so out of hand.[35] We begin with some history.

The term "incel," which was derived from "involuntary celibate," was coined by "Alana Boltwood" (a pseudonym the creator of the term uses to protect her offline identity) in the early 1990s as part of Alana's Involuntary Celibacy Project, an upbeat and inclusive online support group for the romantically challenged.[36] However – as of this writing – "incel" has lost any associations with positivity or inclusiveness.

In the 2000s and early 2010s, Alana's Involuntary Celibacy Project inspired the founding of other incel websites, several of which were dominated by conversations that were "a cocktail of misery and defeatism – all mixed with a strong shot of misogyny."[37] Here are some representative comments from one such site, *Love-Shy.com*:

> The bulk of my anger is over the fact that virtually all women are dishonest to the point that even they themselves believe the lies they tell.
>
> The reality, and I make no apologies for saying this, is that the modern woman is an impossible to please, shallow, superficial creature that is only attracted to shiny things, e.g. looks and money.

By some point in the early 2010s, "incel" and "involuntary celibate" were yoked to the negative, misogynistic thread of inceldom. One turning point was a highly publicized murder spree in Isla Vista in 2014, where Elliot Rodger – a self-identified incel – murdered six college students as part of a "revolution" against women and feminism.[38] Incels are now so strongly associated with misogyny and violence that the Southern Poverty Law center describes them as part of the "online male supremacist ecosystem" and tracks them on the center's "hate map" (a geographical map of hate groups in America).[39]

An online petition that called for banning r/incels described the group as "a toxic echo chamber for its users, [. . .] a dark corner of the internet that finally needs to be addressed."[40] The petition details the problematic content on r/incels:

[34] "Update on Site-Wide Rules Regarding Violent Content."
[35] Of course, it is possible that in this particular case Reddit could not have prevented r/incels from becoming so toxic; perhaps that was inevitable. Nevertheless, we would like to explore some steps Reddit may have taken as a preventative measure.
[36] Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."
[37] Baker, "What Happens to Men Who Can't Have Sex."
[38] Glasstetter, "Shooting Suspect Elliot Rodger's Misogynistic Posts Point to Motive."
[39] Janik, "'I Laugh at the Death of Normies': How Incels Are Celebrating the Toronto Mass Killing."
[40] Cochran, "Shut Down the Subreddit r/incels."

> Violence against women is encouraged. Rape against women is justified and encouraged. [. . .] Users often canonize Elliot Rodger, [. . .] [who is] is often referred to as "Saint Elliot" with many praising his actions and claiming that they would like to follow in his path.[41]

Recall that we have described echo chambers as structures that, like epistemic bubbles, leave relevant sources of information out but, unlike epistemic bubbles, actively discredit those sources.[42] Let us now ask: Was r/incels in fact an echo chamber, as the petition claims?

We think so. r/incels did not just create a space where people with similar ideas about women and feminism congregated. It actively left dissenting voices out. As a petition to have r/incels banned states, "the moderators [of r/incels] have allowed this group to become the epicenter for misogyny on Reddit, *banning users that disagree with the hate speech that floods this forum*."[43]

Dissent was not the only reason users were banned; some were excluded simply for being identified as women. It is difficult to run a rigorous study – it has been noted that "the community [of incels] is deeply hostile to outsiders, particularly researchers and journalists"[44] – but polls have found that r/braincels (another popular incel subreddit, which was banned in 2019) is nearly all men.[45] These demographics were kept up in part through the use of banning; it has been reported that in r/braincels, women were banned "on sight."[46] All the while outsiders (derisively referred to as "normies"[47]) and women ("femoids,"[48] "Stacys,"[49] and "Beckys"[50]) were demeaned in the conversations they were excluded from.

This is disconcerting for many reasons, not least of which is the vulnerability to misinformation about women and dating manifest in users drawn to groups like r/incels. Consider the stories of the pseudonymous "Abe" and "John," each of whom seems to be a typical incel.[51]

41 Cochran.
42 Nguyen, "Echo Chambers and Epistemic Bubbles."
43 Cochran, "Shut Down the Subreddit r/incels" (emphasis added).
44 Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."
45 Beauchamp.
46 Beauchamp.
47 "[A]nyone who is broadly neurotypical, average-looking and of average intelligence." See Squirrel, "A Definitive Guide to Incels."
48 "A portmanteau of 'female' and 'humanoid' or 'android,' this term is used to describe women as subhuman or non-human. Some incels go further and use the term 'Female Humanoid Organism,' or FHO for short." See Sonnad and Squirrell, "The Alt-Right Is Creating Its Own Dialect. Here's the Dictionary."
49 Women considered to be "air-headed, unintelligent, beautiful and promiscuous." See Squirrel, "A Definitive Guide to Incels."
50 "[T]he 'average' woman [. . .] who 'will likely die [sic] her hair green, pink, or blue after attending college' and 'posts provocative pictures because she needs attention' despite being a '6/10.'" See Jennings, "Incels Categorize Women by Personal Style and Attractiveness."
51 The stories of both John and Abe can be found in Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."

Abe, 19, is a lifelong loner who claims to have once dated someone for a month. Abe turned to the internet for support. There he found a cadre of people who were happy to reinforce his belief that the problem was his looks (this is a common incel trope – that our romantic futures are determined by superficial, genetically encoded traits such as height, the strength of one's jawline, and length of one's forehead), and "how manipulative some women can be when seeking validation" (this reflects another trope – that women are shallow, opportunistic, and cruel).[52]

One helpful way to understand this process comes from Alfano, Carter, and Cheong's notion of technological seduction, which we first encountered in the previous chapter.[53] The core idea is that people can encounter ideas that fulfill psychological needs and are consistent with personal dispositions and be attracted – *seduced* – into reading, listening, and watching more related ideas. This often happens in a way that ramps up or becomes more extreme. Users become seduced into a kind of self-radicalization. They need not have had an antecedent belief in the seductive ideas to start identifying with them. The Abe example exhibits these characteristics. He came to the incel community predisposed to think he was an especially bad case and that women were cruel. The community was happy to indulge these thoughts. He then spent more time in the community and began to adopt even more fatalistic views about his dating prospects and more cynical views about women.

John, like Abe, turned to incel groups for support due to feelings of isolation. He too thinks that immutable features of his appearance have doomed him to a life of romantic isolation:

> Most people will not be in my situation, so they can't relate. They can't comprehend someone being so ugly that they can't get a girlfriend [. . .] What I noticed was how similar my situation was to the other guys. I thought I was the only one in the world so inept at dating.[54]

The truth, of course, is that many – if not most – people can relate to these feelings. As Beauchamp notes, "All of us have, at one point, experienced our share of rejection or loneliness."[55] But when socially isolated young men congregate around the idea that they are uniquely a bad case, that anyone who says otherwise is an ideological foe, and when they can exclude perceived ideological foes from their universe of information (as well as anyone whom their conspiracy theories scapegoat and stereotype), the result is a toxic mix of radicalizing ideas and people vulnerable to their uptake. Note that John, too, seems to be a victim of technological seduction.

[52] Beauchamp.
[53] Alfano, Carter, and Cheong, "Technological Seduction and Self-Radicalization."
[54] Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."
[55] Beauchamp.

So what might we do to ameliorate this while respecting the autonomy of the members of groups like r/incels? In what follows we investigate two epistemically paternalistic approaches, one that involves making access to alternative points of view salient, another involves making the barriers of the echo chamber itself more porous.

6.2.1 *Access to Reasons*

Cass Sunstein discusses a number of remedies to echo chambers and filter bubbles that involve providing access to reasons that speak against the ideology of the chamber or bubble.[56]

One such remedy involves the introduction of an "opposing viewpoint button," inspired by an article by Geoffrey Fowler, arguing that Facebook should add a button to News Feed. This button would, in Fowler's words, "turn all the conservative viewpoints that you see liberal, or vice versa."[57] This would enable users to "realize that [their] [. . .] news might look nothing like [their] [. . .] neighbor's."[58] Such a button might not make very much sense in a subreddit, but a variation of it might. Perhaps Reddit could offer dissenting groups the opportunity to have a link posted to a subreddit's menu that would take users to a statement or page that outlines an opposing point of view.

Or, perhaps, instead of a link to a statement, subreddits could have links to deliberative domains, "spaces where people with different views can meet and exchange reasons, and have a chance to understand, as least a bit, the point of view of those who disagree with them."[59]

Whether it is via a link to a statement or a deliberative domain, both proposals involve the adding of an option to access reasons from the opposing side. Were either taken unilaterally and in the spirit of improving users' epistemic lot, they would be instances of epistemic paternalism.

Now we can ask, were Reddit to explore this option of adding access to reasons – either through an opposing viewpoint button or link to a deliberative space – would it be permissible?

We think so, and we will explain why in familiar fashion. The relevant perspective from which to view the intervention is that of the denizens of r/incels. The intervention would not limit the freedom or autonomy of any of the members of the group, nor would it harm them. They are left free to have whatever discussions they please, post whatever they'd like to, and so on. If their minds are changed by the intervention it will be through the cool exchange of reasons, a process of changing their minds from which they cannot be alienated.

[56] Sunstein, *#Republic: Divided Democracy in the Age of Social Media*.
[57] Fowler, "What If Facebook Gave Us an Opposing-Viewpoints Button?"
[58] Fowler.
[59] Sunstein, *#Republic: Divided Democracy in the Age of Social Media*.

6.2.2 *Inclusiveness*

While the proposals under the banner "access to reasons" are promising and permissible, such proposals might not go very far in addressing the issue of echo chambers. It's likely that many users simply wouldn't take advantage of the opportunity to use the links. And, if they did and the experience changed their mind, they would likely leave or be exited from the echo chambers they belonged to. So, while the above proposals may help an individual user escape an echo chamber, the "access to reasons" proposals – on their own – are likely not enough.

Let us, then, explore a proposal that may offer further assistance. At the moment, the moderators of subreddits (and similar entities such as Facebook groups and so on) are free to ban users from their discussions at their discretion. We saw earlier that women were banned from r/braincels "on sight."[60] This, clearly, does not help the community's distorted view of women. The power to ban gives moderators the ability to form and maintain echo chambers, as it gives them the power to literally exclude certain voices from their discussions.

Another class of interventions, then, might be aimed at limiting this power. What might this look like?

Sites like Reddit could give its users some entitlement to not being excluded from subreddits strictly for belonging to a protected class, and this could be accomplished by modifying moderators' privileges to ban users at their own discretion. The site could, for example, discourage discrimination on the basis of protected attributes by stating that it is a behavior that is not allowed on the site and (partially) enforcing this by making some kind of appeals process for bans or suspending moderators who violate the policy.

As a supplement or alternative, a similar system could be set up for bans that do not result from breaking any site-wide or explicitly stated group rules. The idea here is that groups that have moderators who want to ban ideological foes would have to at least do so openly or not at all. The hope here is that many groups would not be okay with this as an explicit policy, reducing or eliminating cases where moderators have an unofficial policy of banning ideological foes.

Anyone familiar with Reddit might object to this suggestion on practical grounds, saying that the site is too large and anonymous for this to work. There are roughly 2 million subreddits in existence, with single subreddits having tens of millions of users.[61] Further, users are typically anonymous, and it is very easy to make new accounts. As a practical matter, the objection goes, such a change to the site is just not feasible.

Reddit's scale and design do present practical difficulties for these proposals, but it does not make them unworkable. There are various ways in which the proposals can be implemented at scale. For example, the site could – for practical reasons – rule that appeals can only be made by certain accounts, such as accounts that have existed for more than a certain amount of time and have been verified. And this rule

[60] Beauchamp, "Incels: A Definition and Investigation into a Dark Internet Corner."
[61] Reddit Metrics, "Top Subreddits."

https://doi.org/10.1017/9781108895057.006 Published online by Cambridge University Press

could, of course, be enforced using algorithms. Penalties for frivolous appeals could also be part of the policy. The site could also consider limiting investigations by only investigating moderators when patterns of appeals appear, for example, once a moderator has racked up a certain number of appeals from verified users in a certain time period. This, too, could be managed algorithmically to make the solution scalable.

Assuming that the practical objection could be addressed, we can then ask: Are these policies permissible?

We think so. To show this, let's look at the policy from the point of view that might have complaints about the proposal: the moderators. What complaints might moderators have? Their autonomy isn't being undercut, nor are they being harmed. So it does not seem that they could make complaints from autonomy or harm. However, they are being constrained in what they can do. So, perhaps, they can complain that their freedom is being encroached upon, specifically in the form of diminished quality of agency. This, it seems, is the only plausible complaint they might have. So it is the one we will explore.

But we now can return to a familiar refrain: That a course of action will limit an agent's freedom is not enough to show that it is wrong. Legal bans on murder limit our freedom, but not wrongfully so. Limiting moderator's privileges to ban users is, we think, a permissible constraint on their freedom. This is because the complaint from moderators that putting a check on banning limits their freedom is complicated by two factors. One factor is that their freedom to ban users limits users' freedom to partake in the conversations they are banned from. So their claim to the freedom to ban users butts up against the freedom of the users they will ban. The other factor is that while some bans are non-objectionable – bans made in response to violations of Reddit's site-wide ban on involuntary pornography, for example – the class of bans we are discussing here is objectionable. Users who have been banned based on their membership to a protected class can reasonably object to those bans and to a system that allows them.

6.3 CONCLUSION

Since much of the internet information environment is epistemically noxious, there is lots of room and opportunity for epistemically paternalistic interventions such as Click-Gap, opposing viewpoint buttons, and modifications to moderators' privileges. Hence, many epistemically paternalistic policies can (and should) be a perennial part of the internet information environment. What should we conclude from that? One thing is that we should recognize that developers should engage in epistemic paternalism as a matter of course. Another is that our focus in evaluating epistemically relevant interventions should not be on whether such actions are epistemically paternalistic. Rather, it should be on how they relate to other values (such as well-being, autonomy, freedom, and so on).