



HOKKAIDO UNIVERSITY

Title	Social-Media Companies Have a Duty to Remove Misinformation
Author(s)	Campbell, Douglas R.
Citation	Journal of Applied Ethics and Philosophy, 17, 9-17
Issue Date	2026
DOI	https://doi.org/10.14943/jaep.17.9
Doc URL	https://hdl.handle.net/2115/99135
Type	departmental bulletin paper
File Information	JAEP17_9-17.pdf



Social-Media Companies Have a Duty to Remove Misinformation

Douglas R. Campbell

Department of Philosophy, Alma College

Abstract

The goal of this article is to argue that social-media companies ought to remove misinformation from their platforms. I begin by outlining four kinds of harms that misinformation can produce. In the following section, I clarify the conceptual foundations of this duty, including the difference between purposeful and general social-media sites, influenced by Lavinia Marin's recent research, and the correct characterization of online debates about topics such as climate change and vaccination.

Keywords: social media; misinformation; choice architecture; epistemic harm; disinformation; infodemic; information environment

I argue that social-media companies have a duty to remove misinformation. This duty exists in order to mitigate four kinds of harm: harm to a person's well-being, harm to democratic institutions, harm to an information environment, and the harms that result from a pernicious choice architecture. For the purposes of this article, it is sufficient to define misinformation as content that is either false or misleading.¹ This capacious understanding of misinformation includes content that asserts something false, content that asserts something true but omits important facts in tendentious ways, and deepfakes, which are not, strictly speaking, true or false but which are misleading. I argue that such content is harmful in the first section. However, most of this article is not about why such content is harmful; instead, I spend most of this article walking through important conceptual foundations that underlie the duty to remove misinformation. For instance, Lavinia Marin's distinction between purposeful and general social-media sites is key here, and so is Ludwig Wittgenstein's understanding of a language-game.

Social-media companies occupy such an important role in 21st-century society that they have a duty to remove misinformation. They should prioritize removing content in proportion to how harmful it is.

Accordingly, content that undermines people's health and well-being and content that undermines democratic institutions should be removed first. Content that, in contrast, attributes incorrect titles to Taylor Swift's albums is not much of a priority, even though the false beliefs generated by that misinformation do harm our information environment.

As I argue below, the common defense that removing misinformation is tantamount to censorship and violates free-speech rights should not make us doubt the duty to remove misinformation. In the first place, such a defense implausibly generates a duty to not remove misinformation. More importantly, it ignores the place that social-media companies occupy in today's society, and it is based on a characterization of online discourse that I argue is false.

1. Harms of misinformation

A duty to remove misinformation amounts to a duty to mitigate the harms caused by social-media sites. In this section, I shall identify four harms that misinformation causes and that the duty to remove misinformation would mitigate. The harmful nature of misinformation has been subject to fruitful and thorough research in recent years, and I shall sometimes use the footnotes to direct readers

¹ See Carey 2023 for a more sophisticated treatment of misinformation as content that has a disposition to worsen a person's overall epistemic position.

to thoughtful discussions elsewhere. It is not the goal of this article to contribute to this research exactly, but to justify the claim that there exists a duty to address the harms caused by misinformation. For this reason, I shall move briskly through some of the harms, saying enough to establish that the harms exist but not much more.

The first kind of harm generated by misinformation is the simplest: the way that false beliefs make us worse off as individuals. This is most evident in the case of misinformation that concerns our health and well-being. For example, the misinformation that proliferated online during the COVID-19 pandemic harmed people by persuading them that getting vaccinated was dangerous. Social-media platforms give airtime to people such as the so-called Disinformation Dozen, a small group of online influencers who were responsible for roughly two-thirds of all anti-vaccine misinformation during the pandemic.² One member of the Disinformation Dozen, Sherri Tenpenny, gained such a high level of prominence on account of her social-media presence that she eventually testified in front of the Ohio House of Representatives' health committee. Her testimony reflected the beliefs that she spread online, namely, that the COVID-19 vaccine would magnetize people and that keys and other household items would stick to people who were vaccinated. Early in 2022, she predicted that "by the end of 2022, every fully vaccinated person over the age of 30 may have the equivalent of full-blown vaccine-induced immune suppressed AIDS."³ Such misinformation can convince people to avoid important medical interventions.

There is nothing special about COVID-19 misinformation in this regard. Social-media sites have been hotbeds for pro-anorexia communities for a long while.⁴ People in these communities convince others not to get treatment for their psychiatric condition, which is terribly unfortunate in light of the fact that anorexia is more lethal than any other mental illness. This misinformation can have obvious and measurable negative effects on a person's life, even to the point of death.

Health-related misinformation stands out here as the most obvious culprit, but other forms of misinformation can harm a person's well-being, too. People in so-called incel, or involuntarily celibate, communities can become so entrenched in misogynistic and sexist misinformation that they miss out on meaningful relationships with women. Members of QAnon and flat-Earther communities can similarly become so obsessed with the misinformation and conspiracy theories that circulate around them that they become estranged from people in

their lives. That too is harmful.

The second kind of harm is the damage done to democratic institutions. The democratic process works in the following manner: I, a voter, have values; competing politicians express values and support policies based on those values; I vote for the politician who most closely reflects my own values and consequent policy preferences; the elected official then works in the government alongside other elected officials to draft policies that reflect those values. The democratic process breaks down under the weight of misinformation. Former president of the U.S.A., Barack Obama, got at this point well when he said:

If we do not have the capacity to distinguish what's true from what's false, then by definition the marketplace of ideas doesn't work. And by definition our democracy doesn't work. We are entering into an epistemological crisis.⁵

To understand why this is the case, imagine a candidate for political office spreading the misinformation that climate change is a hoax perpetrated by the Chinese government.⁶ There are many voters who care about the environment and who deeply value ensuring that future generations inherit a healthy planet. Those voters might be persuaded by this candidate that climate change simply does not exist, and then they vote accordingly. The policies that result from this elected official will reflect the belief that climate change is not occurring, even though it is, and will therefore fail to align with the values of the electorate (even though it aligns with their false beliefs). Misinformation threatens the integrity of democratic processes by thwarting the representation of voters in policy-making.

The third kind of harm is the poisoning of our information environment.⁷ The more false or misleading content there is online, the worse our position with respect to the formation of beliefs becomes. We will tend to encounter more false or misleading assertions, in general. The quality of the average piece of information declines because there is a greater quantity of low-quality information available. High-quality sources of information might be falsely slandered, which in turn will cause us to miss out on true beliefs that those sources of information might have provided us with. We could very well even abandon true beliefs that we had already formed through testimony provided by high-quality sources; the slander against them could incorrectly convince us to drop those beliefs. Online influencers trade in misinformation all too frequently: the Disinformation Dozen, mentioned earlier, is one

² See Srikanth 2021.

³ See Gage 2022.

⁴ Parsell 2008 has an interesting discussion about why pro-anorexia and similar communities thrive online.

⁵ Goldberg 2020.

⁶ See Emery 2016.

⁷ This is a complicated kind of harm, which I unpack only in a cursory manner here. See Carey 2023, de Ridder 2021, and Marin 2021b for more thorough discussions.

example. In this information environment, going online is akin to wading through vast quantities of misinformation. Even the vigilance required of us as doxastic agents is exhausting, which is a kind of harm in itself.

The last kind of harm is the harm produced by a choice architecture that is built on misinformation. This kind of harm is rarely, if ever, mentioned in discussions of misinformation, and it is subtle. The term ‘choice architecture’ originated in libertarian-paternalist research.⁸ The choice architecture is the context in which a choice is made, and libertarian paternalists view modifications of the choice architecture that they call *nudges* as permissible ways of promoting the likelihood that some particular choice will be made, without the use of coercion. A clear offline example is the design of grocery stores: customers are being nudged towards the purchase of candies by placing them prominently near cash registers, where customers generally spend a few minutes waiting. The likelihood that customers will buy candies increases, but no coercion is used. Social-media developers design the choice architecture of their sites in ways that maximize the amount of time spent on them.⁹ For instance, notification icons are designed to be a shade of attention-catching red, and many video-streaming sites play the next video automatically after the completion of the current one.

Misinformation plays a key role in the choice architecture of a social-media site. One way to get people to spend more time on the site is to show them content that will grab their attention. One prominent kind of nudge is the recommendation that a site will offer by saying that some content is trending or worth checking out. The problem is that sometimes this content is misinformation because the algorithm or the developers have predicted that the content that will be most effective at maximizing user-engagement metrics is misinformation. Certainly, one of the most heinous examples of this was Meta’s knowing promotion of pro-eating-disorder content on Instagram in order to nudge young girls to stay on the platform longer.¹⁰ The removal of misinformation would clean up the choice architecture of a site, no longer allowing developers to promote clickbait or other misleading or false content in order to boost engagement metrics.

If user-engagement metrics fall as a result of the removal of misinformation, that might well be a good thing. There are randomized controlled trials showing that reducing social-media usage improves anxiety, fear of missing out, loneliness, and depression.¹¹ If

changes to the choice architecture entail that we spend less time online, that might be better for us. Even if this psychological research turns out to be false upon further investigation, many philosophers have argued that we have a right to our attention that the constant distractions posed by social-media sites violate.¹² The duty to remove misinformation would not stop people from distracting us with the truth, but it would remove many of the usual weapons, such as clickbait headlines, that sites use to distract us from whatever else we would prefer to spend our time on.

There is much to be said about the way that social-media platforms’ algorithm plays a role in aggravating the harms of misinformation. Since their algorithms are proprietary and thus not available for our scrutiny, it is hard to speak about them in concrete terms. Yet, we can speak generally about this problem. The algorithmic boosting of some content for the sake of user-engagement metrics amplifies misinformation and increases the harm done. In a profound sense, social-media sites do the opposite of what this article argues that they ought to do: instead of removing, they promote misinformation. Even if they changed their algorithms to downplay misinformation, it would still be wrong of them to fail to remove misinformation because the harms outlined in this section would still occur, even if at a reduced rate. For this reason, social-media sites ought to outright remove misinformation.

Ultimately, these are four kinds of harm that the duty to remove misinformation would mitigate. A careful bit of reflection reveals that not all misinformation is equally harmful. Misinformation that threatens our well-being and democratic institutions is the most harmful. Health-related misinformation can cause the most visible harm, whereas content that undermines democracies is most likely the most consequential per unit. However, even apparently harmless misinformation, such as incorrectly stating that Taylor Swift was born in 1988, worsens our information environment and is thus responsible for at least one of the four harms. In what follows, I shall present the conceptual foundations of the duty to remove misinformation, and I maintain that social-media companies should remove content while prioritizing what is most harmful. Misinformation about Taylor Swift, and similar subjects, should be removed last, and the most harmful misinformation should be removed first. Ultimately, all misinformation should be removed because even the least harmful kind still produces harm to our information environment, but some misinformation produces even more harm and, consequently, should be prioritized.

⁸ See, e.g., Thaler and Sunstein 2008 and Campbell 2022.

⁹ Campbell 2022 discusses many examples of such nudges as well as their moral status.

¹⁰ See Wells, Horwitz, and Seetharaman 2021.

¹¹ E.g., Hunt et al. 2018.

¹² E.g., Puri 2021, Tran 2015, and Chomanski 2023.

2. The conceptual foundations of the duty to remove misinformation

We should remove misinformation in order to prevent or mitigate these harms. However, there is a problem. The most prominent objection to removing misinformation is that doing so constitutes censorship and violates people's free-speech rights. This is built on a way of setting up content-moderation dilemmas as a competition between two values: respecting freedom of speech and mitigating harm. This way of setting up the problem is common both inside and outside academia.¹³ There is an interpretation according to which this objection is simply false. People have free-speech rights to protect them from government overreach. To say that a company's decision to remove your misinformation violates your free-speech rights is as false as saying that your mother's decision to tell you to stop talking violates your free-speech rights. Private individuals are morally permitted to remove content from their platforms. This is how we might reply to this interpretation of the free-speech objection. However, there are two much more compelling and intellectually fruitful ways of construing this objection.

First, we can establish a more compelling version of this objection by reframing it as being less about people's free-speech rights and more about the integrity of public discourse on social-media sites. Sites such as X and Instagram occupy such an important role in democratic

societies that they might be plausibly conceived of as the spaces in which citizens work out ideas together. The elimination of an important idea, even a false one, such as that climate change is a hoax perpetrated by a foreign government, from social-media sites effectively eliminates it from public discourse altogether. It is not fair to point out that people in free societies could still circulate such ideas in print, by means of brochures and pamphlets: ideas circulated on YouTube, for example, reach an audience with a speed and scope of distribution that is not paralleled by any other form of communication that humanity has ever developed. The removal of even false ideas harms public discourse by taking out of the public's consideration possibly important things to reflect on and debate.

This objection to the duty to remove misinformation is powerful, but I disagree with it. It is a helpful objection to consider because it provides us with an opportunity to reflect on the conceptual foundations of the duty to remove misinformation.

We should begin by observing exactly what the objection entails. Specifically, it entails that there is a duty to *not* remove content on social-media sites. This is not merely the negation of the central claim of this article, which is that there is a duty to remove misinformation. The objection affirms that it is morally impermissible to remove content posted online. The duty to not remove content is implausibly strict. Imagine Asian owners of a social-media site. Now imagine that on this site, many users post misinformation that claims that Asian people uniquely spread some disease and that anti-Asian racism is, therefore, justified. This is not unlike what happened at times throughout the COVID-19 pandemic amid speculation concerning the origins of the virus. If the owners of the site had a duty to not remove this misinformation, then they would not be permitted to remove any of this content. It would not matter how much it violated their own values or how uncomfortable they felt owning a site that was responsible for airing so much racist content. People would have to tolerate misinformation on their own platforms even when they had the tools and desire to remove it. If the value of their company started to decrease rapidly, they would still have to tolerate it.

Consider how counterproductive this would be. One important tool for social change is the pressure that business owners, and other similarly positioned people, can exert on people with socially undesirable behaviors, such as spreading racist misinformation. If we have a duty to not remove misinformation, then we lose this important tool for social change.

Moreover, social-media sites *already* remove all sorts of content, and it is obviously good that they do so. The removal of child pornography is a clear example of content that we already recognize ought to be removed.

¹³ For a recent academic example, see Kozyreva 2023. See also Stroud 2019's discussion of what he calls trade-offs in the attempt to remove misinformation. Here are a few examples from outside academia. First, Monbiot 2021's way of setting up the problem: "we have a right to speak freely. We also have a right to life. When malicious disinformation — claims that are known to be both false and dangerous — can spread without restraint, these two values collide head-on." Second, see Zuckerberg 2019's discussion about the practices of his own company: "we make a lot of decisions that affect people's ability to speak. [...] Frankly, I don't think we should be making so many important decisions about speech on our own either." This is reflected in Mudde 2020's headline from *The Guardian* during the pandemic: "don't let free speech be a casualty of coronavirus. We need it more than ever." Howard 2021: 168 sums up this debate thus: "that something should be done to confront this misinformation seems undeniable. The controversy arises when we consider what, exactly, the appropriate solution is. One possibility is to reduce the circulation of misinformation by insisting that social media networks remove it from their platforms, or at least reduce its dissemination. Yet in the public debate, this suggestion continually inspires a chorus of opposition from those who think that such a move would run afoul of one of the most cherished values of liberal democracy: freedom of speech" (emphasis in original).

So, clearly, there is no duty to *not* remove content in general. As for misinformation in particular, sites already have unobjectionable removal practices. For instance, imagine if I post a false review of a restaurant on a review site such as Google Reviews or Yelp. Imagine that I do this specifically in order to skew perceptions of the restaurant by remarking on dishes that do not exist there: I talk about chicken tikka masala and naan in a review for an Italian restaurant. Google and Yelp would delete my review according to their present removal guidelines, and nobody would disagree with this removal. It is clear that such misinformation's pollution of our information environment cannot and should not be tolerated. The whole purpose of the review site falls apart if it tolerates false reviews.

I said earlier that there were two intellectually fruitful ways of construing the free-speech objection. The first, we have seen, entails that there is a duty to *not* remove content, and I have argued that, on the basis of this entailment, this objection does not work. The second denies that there is a duty to *not* remove content; instead, it maintains that it is permissible to leave content untouched. It is permissible to remove, but it is permissible to *not* remove, too. This objection is plausible to the extent that it avoids what was unintuitive about a duty to not remove, but it overlooks the tremendous harm that misinformation can cause, which I discussed in the first section. The duty to remove content online *is* a duty to mitigate harm. Since this objection *concedes* that there is no duty to *not* remove content and that removing content does not trample on free-speech rights of users, it is mysterious why failing to mitigate the harms done by misinformation is permissible.

The need to remove misinformation is even greater when we consider cases that bear on public discourse about weighty matters. Imagine if someone spreads misinformation concerning the backgrounds and political views of some candidates for political office on Wikipedia. Wikipedia is so important to many people's research into candidates that the need to remove this information increases in urgency. This is partly why the objection that highlights the place of social-media platforms in 21st-century society fails: *the place of social-media platforms is so important to public discourse that the duty to remove misinformation follows naturally*. The objection drew the wrong conclusion from the importance of social-media platforms: it concluded that we should let discussions unfold without intervention, but, in fact, interventions are needed because of how important they are.

I have cited hypothetical cases on Wikipedia, Google Reviews, and Yelp. At first, these might seem disanalogous to cases on Facebook, Instagram, TikTok, and so on. That is because the former are *purposeful* social-media sites, whereas the latter are *general*

social-media sites. The distinction between purposeful and general sites was invented by Lavinia Marin in a discussion that clarified the norms of sharing information online.¹⁴ Purposeful social-media sites are those that have a clear purpose that unifies the behaviors of users on the site, and this purpose gives us norms that we can use to assess whether someone's behaviors are objectionable. For instance, LinkedIn is dedicated to discussions about users' careers. If someone is spreading misinformation about his or her employment history, then it is obvious that such content should be removed. We can arrive at that judgment by citing the site's purpose. This is what is going on when we agree that false, misleading reviews on Yelp should be removed.

It is trickier to assess cases on general social-media sites because they lack a clear, unifying purpose. That is what it means to be a *general* site: users use such sites as Facebook and Instagram for a variety of purposes, perhaps even changing purposes of use daily. The fact that Facebook, Instagram, Reddit, and so on, allow such a variety of uses that users find acceptable makes it harder for us to justify the removal of content. Some Reddit communities might be dedicated to finding a romantic partner, in which case spreading misinformation about oneself is wrong and deceptive, whereas there are other Reddit communities that might be satirical, in which it would be uncharitable to view posts as deceptive and misinformation.

Nevertheless, I maintain that we can still justify the removal of misinformation from general social-media sites. Firstly, we could still point to the harms of misinformation, enumerated in the first section of this article, to justify the duty to remove misinformation on simple consequentialist grounds. However, I think that there is an important philosophical concept that clarifies the matter. This is the concept of a *language-game*, which was invented by Ludwig Wittgenstein and introduced into the ethics of social media by Lavinia Marin.¹⁵

Wittgenstein proposed the concept of a language-game in order to explain that our words and utterances get their meanings from the contexts in which they are embedded. Here is an example that he gives:

The language is meant to serve for communication between a builder A and an assistant B. A is building with building stones: there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words "block," "pillar," "slab," "beam." A calls them out; — B brings the stone which he has learnt to bring at such-and-such a call. — Conceive this as a complete primitive

14 See Marin 2021a.

15 See Wittgenstein 1958 and Marin 2021a.

language.¹⁶

This simple language-game suffices to illustrate the point. After all, shouting ‘block!’ could mean different things in different language-games. Wittgenstein’s concept is rich and deep, but for our purposes here, we can focus on the way that utterances are considered acceptable or illegitimate depending on the context in which they are uttered. For instance, poetry is an entirely different language-game than a broadcast on CNN, and different utterances are acceptable in each. I emphasize the centrality of Wittgenstein’s concept here because it will prove indispensable to our analysis of misinformation on social-media sites.

Now let us return to the distinction between purposeful and general social-media sites. Each purposeful social-media site is a different language-game. There are different norms for communicating about ourselves on LinkedIn than there are on Tinder or Bumble. There are things that we might say about ourselves on Bumble that would raise eyebrows on LinkedIn, and a conversation focused on our professional history is par for the course on LinkedIn but would be entirely out of place on Tinder. While each purposeful site has its own distinct language-game, the rules are easily discerned. To figure out what is acceptable to share on LinkedIn is, in the vast majority of cases, as easy as it is to figure out that if someone at a construction site says the word ‘block!’ to you, he or she is asking for a block.

General social-media sites look different because of the variety of uses. There is *some* language-game happening on Facebook, but it is not always clear what the language-game is. Consequently, the rules are not so obvious; neither is what is acceptable. However, we are armed with some psychological and sociological research that illuminates the answer to this question. People generally post on social-media sites for entertainment, as part of status-seeking, for self-expression, or for socialization.¹⁷ Some researchers say that the diversity of motivations can be reduced to two basic needs: the need to belong and the need to present oneself to others.¹⁸ One group of researchers summarizes their findings by saying that “catchiness rather than truthfulness often drives information (and misinformation) diffusion on social media.”¹⁹ The language-game on general social-media sites is something like the language-game on a playground: the kids talk about a variety of subjects, but they are not really interested in the truth, by and large.²⁰

They speak, spread rumors, and share gossip in order to fit in, show what their values are, have fun, exclude people whom they want to exclude, and promote social cohesion among people that they want to include.

In light of this research, we should not have any objections to removing misinformation. The objection to the duty to remove misinformation was founded on the view that conversations on social-media sites, such as X, were part of an important public debate that it would be wrong to intervene in. On the one hand, when such a debate is happening, we *ought* to intervene for the same reason that we think that Wikipedia’s content should be carefully scrutinized: the cost of negligence and laxity is too high. On the other hand, research by psychologists and sociologists suggests that conversations on social-media sites should not generally be characterized this way. We should be as ready to jump in and intervene in these conversations as we would be if we overheard kids on a playground spreading rumors and gossip. After all, as bad as the children’s rumors and gossip might be, the misinformation spread on social-media sites can be much more dangerous.

It is because of this increased danger that we should respond differently to misinformation online than we would to rumor-spreading on a playground. Whereas we might respond to the spread of a rumor by privately correcting the offending child, the harms caused by social-media platforms instead call for the removal of the content. The harms I cataloged in the first section of this article generate this duty for us, but to differing degrees. The one most salient, for the topic at hand, is the harm done to our information environment. The *only* way to clean up our information environment is to remove the content. Otherwise, we will still have a polluted information environment.

There is an objection to this view that conversations on social-media sites are like conversations on a playground. The objection is that although research has revealed that this is how people *post* online, we have reason to believe that people *consume* content differently. Around half of Americans get news from social-media sites.²¹ The spread of misinformation on social-media sites is, to some extent at least, due to an unfortunate asymmetry: *people post as if they were in a playground, but they consume as if there were a serious debate occurring between knowledgeable people.*

Let us pursue a bit more the possibility that conversations on social-media sites are important and high-minded debates that are a worthy source of news information. This view is not plausible in light of what we know about the motivations of those who post, but it is worth taking seriously since many people consume news on social-media sites as if it were true. It might be the case that this view entails that we should be reluctant

16 Philosophical Investigations §2. Translated by G.E.M. Anscombe.

17 See Chen et al. 2015, Diddi and LaRose 2006, and Lee and Ma 2012.

18 See Nadkarni and Hofman 2012.

19 Chen et al. 2015: 111.

20 Marin 2021a similarly infers that conversations on general social-media sites are like the gossip and rumors spread by kids on a playground.

21 See Walker and Matsa 2021.

to remove ideas from these debates. However, I think that we should ultimately reject this possibility because these sites are simply *bad venues for hosting the sorts of debates that are characterized this way*. Consider some of the misinformation that gets spread online regarding climate change or vaccination. Recall Sherri Tenpenny's misinformation, mentioned earlier, that the COVID-19 vaccine magnetizes people. These are important scientific claims. They should be tested in whatever experimental setting is appropriate: perhaps randomized controlled trials, or something else, where appropriate. The proper institutional venues for putting forward these claims are academic ones, with the infrastructure provided by peer review. There might be flaws to institutional peer-review systems, but none of these flaws are remedied by Instagram or YouTube. Social-media sites, in turn, offer precisely *zero* advantages with respect to reliability or accuracy.

In other words, psychological and sociological research makes it clear to us that people are *not* posting for reasons that would make us refrain from removing content online. People do generally consume content as if there were important debates happening on social-media sites, but to the extent that people sincerely believe that this is happening, they are mistaken. The important debates concerning the topics that misinformation is often about are happening elsewhere, in settings that have reliable, even if imperfect, institutions guiding them. Moreover, *if* there were truly important debates happening on social-media sites, the importance of those debates would entail that we ought to intervene in order to ensure the integrity of the debates, just as we intuitively recognize the importance of ensuring that Wikipedia's information is accurate.

Without valuable institutional safeguards, debates on social-media sites are insensitive to the truth in ways that were evident as soon as we learned the motivations that people generally have when posting. Some people might still insist that, even without institutional peer-review processes, high-quality content will still rise to the top. This objection goes like this: perhaps peer review has merits, but we can get by without it in other settings and still arrive at high-quality information. I present one experiment in order to rebut this objection. Researchers performed an experiment on Reddit in which they made a comment and then either upvoted it immediately, downvoted it immediately, or left it alone.²² If high-quality content rises to the top, then we would expect that, over time, the initial upvote or downvote would be canceled out by corrections from other users. In fact, that is not what happened at all. Both the upvote and the downvote had the effect of increasing overall turnout:

the initial popularity of the comment led to sustained popularity in the long term. The upvote predicted further upvotes down the road. In other words, people like content that other people like, even when the initial liking is manufactured and artificial. Interestingly, in the experiment, while the downvote increased turnout, it did not predict further downvotes. Overall, then, we should conclude from this research that we have no reason to believe that the success of content online tracks anything related to the truth. Perhaps the success of content is due to luck or the money that some influencers spend buying followers or employing bots, or some other factor.²³ This insensitivity to the truth provides us with another reason to abandon any remaining reluctance to remove misinformation from social-media sites.

Lastly, I will concede that it can be hard to tell what counts as misinformation. I do not mean that the definition of misinformation is unclear. I mean that it is sometimes difficult to determine whether some piece of online content is an instance of misinformation. For instance, in the early days of the COVID-19 pandemic, content that concerned the efficacy of masking or the possibility of asymptomatic transmission of COVID-19 was rife on social-media sites, but it was not obvious what the truth was. This point is well-taken, but this should not make us doubt the existence of a duty to remove misinformation. It should make us carefully inquire into whether some content is misinformation. We should not lose sight of the fact that misinformation is very often easily identified as such, even if not always.

Moreover, it would be wrong to doubt the existence of a duty on account of the fact that there are sometimes cases in which it is not obvious whether the duty applies. Think about the duty to rescue someone from drowning when you are easily able to do so. Now imagine that you are on a beach and that you see someone who *might* be drowning. Some of the time, it is obvious whether a person is struggling in the ocean, but this is one of those cases in which it is not so obvious. There are many reasonable ways to respond to this: you can try to get a better view of the person, while bearing in mind that time might be sensitive, or you can rush out there and err on the side of caution. Either way, nothing about the difficulty of determining whether the duty applies to this specific case should make us doubt the existence of the duty. The same goes for the duty to remove misinformation and the difficulty of determining whether

22 See Muchnik et al. 2013 for the original research; see also Sunstein 2017: 112 for a discussion.

23 For this reason, I in general think that we should reject the claim that if we subject ideas to competition in a so-called marketplace of ideas, the winner of that competition will be the truth. There is a discussion of this claim in Kozyreva et al. 2022. As Sunstein 2021: 20 puts it, "far from being the best test of truth, the marketplace ensures that many people accept falsehoods."

something *is* misinformation. I recognize that this case of the drowning person is not identical in every respect to the case of misinformation, but the general point is that the existence of a duty should not be doubted on the grounds that it is not always obvious when that duty applies.

A final objection: perhaps misinformation should not be removed; instead, social-media companies should simply flag misinformation as false or misleading. If this is sufficient, then we can preserve the freedom of speech of the users who post the misinformation, while also disarming their content's ability to harm people. There are a few problems with this suggestion, and they are all related to people's cognition. First, the presence of misinformation is polluting the information environment, which in turn harms users' ability to reliably form beliefs well. Therefore, since people are not able to reason well, they are not able to deliberate properly about whether to believe the misinformation or the flag alerting them to the fact that it is misinformation. The flag-but-do-not-remove proposal misunderstands the problem, which, in reality, concerns people's doxastic abilities being undermined. Second, the proposal overlooks the role of cognitive biases, such as confirmation biases, in determining what content people online take seriously. Perhaps some users will be persuaded by the flag, but not all will be. Accordingly, the misinformation should be removed for the sake of those unconvinced users. Third, the flag is no different in principle from everyone else on social-media sites who speaks out against misinformation. If those other people will not make a difference, then neither will the flag. Perhaps users who debunk misinformation are not so visible to users who consume misinformation; so, the hope is that the flag will be seen by the users who are consuming misinformation because it is adjacent to the content. However, the first two problems with this proposal explain why we should not be so optimistic.

The flag-but-do-not-remove proposal claims to resolve the harms of misinformation by merely *adding more speech* to our information environment, but the addition of more speech will not clear up the information environment. It merely contributes to it. The fact that social-media sites already contain accessible, high-quality information does not reduce the damage done to our information environment. So, adding more speech, in the form of flags, will not, either.

For these reasons, I conclude that social-media companies have a duty to remove misinformation.

Bibliography

- Campbell, Douglas R. 2022. "Nudging and Social Media: The Choice Architecture of Online Life," *Giornale Critico di Storia Delle Idee* 2002 (2): 93-114.
- Carey, Brandon. 2023. "Misinformation and Epistemic Harm," *Social Philosophy Today* 39: 89-100.
- Chen, X., Sin, S. C. J., Theng, Y. L., Lee, C. S. 2015. "Why do social media users share misinformation?" In PL Bogen, S Allard, H Mercer, M Beck, SJ Cunningham, D Goh, G Henry (eds.) *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15*, New York, 111-114.
- Chomanski, Bartłomiej. 2023. "Mental Integrity in the Attention Economy: in Search of the Right to Attention," *Neuroethics* 16: 8-19.
- De Ridder, Jerome. 2021. "What's So Bad about Misinformation?" *Inquiry* 1-23 <https://doi.org/10.1080/0020174X.2021.2002187>
- Diddi, A., & LaRose, R. 2006. "Getting hooked on news: Uses and gratifications and the formation of news habits among college students in an internet environment." *Journal of Broadcasting and Electronic Media* 50 (2): 193-210.
- Emery, David. 2016. "Did Donald Trump Claim Global Warming Is a Hoax?" September 27, 2016 <https://www.snopes.com/fact-check/donald-trump-global-warming-hoax/>
- Gage, Brandon. 2022. "Anti-vax activist claims COVID-19 vaccines cause AIDS," *Salon* April 3, 2022. https://www.salon.com/2022/04/03/anti-vax-activist-claims-19-vaccines-cause-aids_partner/
- Goldberg, Jeffrey. 2020. "Why Obama Fears for Our Democracy." *The Atlantic*, November 16, 2020. <https://www.theatlantic.com/ideas/archive/2020/11/why-obama-fears-for-our-democracy/617087/>
- Howard, Jeffrey. 2021. "Coronavirus Misinformation, Social Media, and Freedom of Speech," F. Niker and A. Bhattacharya, *Political Philosophy in a Pandemic: Routes to a More Just Future*. New York: Bloomsbury, 167-176.
- Hunt, Melissa G., Rachel Marx, Courtney Lipson and Jordyn Young. 2018. "No More FOMO: Limiting Social Media Decreases Loneliness and Depression," *Journal of Social and Clinical Psychology*, 37 10: 751-768.
- Kozyreva, Anastasia, Herzog, Stefan M., Lewandowsky, Stephan, Hertwig, Ralph, Lorenz-Spreen, Philipp, Leiser, Mark, and Reifler, Jason. 2023. "Resolving content moderation dilemmas between free speech and harmful misinformation," *Proceedings of the National Academy of the Sciences*. 120: 7 doi: 10.1073/pnas.2210666120.
- Lee, C. S., and Ma, L. 2012. "News sharing in social media: The effect of gratifications and prior experience." *Computers in Human Behavior* 28 (2): 331-339.
- Marin, Lavinia. 2021a. "Sharing (mis) information on social networking sites. An exploration of the norms for distributing content authored by others," *Ethics and Information Technology* 23 (3): 363-372.
- . 2021b. "Three contextual dimensions of information on social media: lessons learned from the COVID-19 infodemic," *Ethics and Information Technology* 23: 79-86.
- Monbiot, George. 2021. "Covid lies cost lives – we have a duty to clamp down on them" *The Guardian* January 27, 2021

- <https://www.theguardian.com/commentisfree/2021/jan/27/covid-lies-cost-lives-right-clamp-down-misinformation>
- Mudde, Cas. 2020. "Don't let free speech be a casualty of coronavirus. We need it more than ever," *The Guardian*, April 6, 2020. <https://www.theguardian.com/commentisfree/2020/apr/06/coronavirus-free-speechhungary-fake-news>
- Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. "Social Influence Bias: A Randomized Experiment," *Science* 341 (6146): 647–51.
- Nadkarni, A., and Hofmann, S. G. 2012. "Why do people use Facebook?" *Personality and individual differences*, 52 3: 243–249.
- Parsell, Mitch. 2008. "Pernicious Virtual Communities: Identity, Polarisation, and the Web 2.0," *Ethics and Information Technology* 10: 41–56.
- Puri, Anuj. 2021. "The Right to Attentional Privacy," *Rutgers Law Record* 48 (1): 206–221.
- Srikanth, Anagha. 2021. "12 prominent people opposed to vaccines are responsible for two-thirds of anti-vaccine content online: report" *The Hill* March 21, 2021. <https://thehill.com/changing-america/well-being/prevention-cures/544712-twelve-anti-vaxxers-are-responsible-for-two/>
- Stroud, Scott. 2019. "Pragmatist Media Ethics and the Challenges of Fake News," *Journal of Media Ethics* 34 (4): 178–192.
- Sunstein, Cass. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton, New Jersey: Princeton University Press.
- . 2021. *Liars: Falsehoods and Free Speech in an Age of Deception*. Oxford: Oxford University Press.
- Thaler, R. and Sunstein, C. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven CT: Yale University Press.
- Tran, Jasper L. 2015. "The Right to Attention," *Indiana Law Journal* 91: 1023–1064.
- Walker, Mason, and Matsa, Katerina Eva. 2021. "News Consumption Across Social Media in 2021," *Pew Research Center* September 20, 2021 <https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/>
- Wells, Georgia, Jeff Horwitz, and Deepa Seetharaman, 2021. "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," *Wall Street Journal* September 14, 2021. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7&mod=article_inline
- Wittgenstein, Ludwig. 1958. *Philosophical Investigations*. Oxford: Blackwell Publishing. Translation by G.E.M. Anscombe.
- Zuckerberg, Mark. 2019. "Standing for voice and free expression," *Facebook* October 17, 2019 <https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/>