Topoi
https://doi.org/10.1007/s11245-024-10101-3

Social Evidence Tampering and the Epistemology of Content Moderation

Keith Raymond Harris[1]

Accepted: 27 August 2024

Abstract
Social media misinformation is widely thought to pose a host of threats to the acquisition of knowledge. One response to these threats is to remove misleading information from social media and to de-platform those who spread it. While content moderation of this sort has been criticized on various grounds—including potential incompatibility with free expression—the epistemic case for the removal of misinformation from social media has received little scrutiny. Here, I provide an overview of some costs and benefits of the removal of misinformation from social media. On the one hand, removing misinformation from social media can promote knowledge acquisition by removing misleading evidence from online social epistemic environments. On the other hand, such removals require the exercise of power over evidence by content moderators. As I argue, such exercises of power can encourage suspicions on the part of social media users and can compromise the force of the evidence possessed by such users. For these reasons, the removal of misinformation from social media poses its own threats to knowledge.

Keywords Content moderation · Dissent · Skepticism · Social media · Testimony · Warrant

1 Introduction

Misinformation and disinformation pose a range of challenges. Some of these are epistemic. The existence of misinformation and disinformation, and perhaps even the realistic possibility of their existence in the social epistemic environment, threaten the acquisition of knowledge, understanding, and other epistemic goods. Some are social. To the extent that misinformation and disinformation compromise the ability of individuals to converge on a shared understanding of certain basic facts, these phenomena can promote polarization and otherwise tear at the social fabric. And some challenges are political. By distorting individual and collective judgments, misinformation and disinformation reduce the quality of voters' decisions and ultimately compromise the abilities of political systems to respond to serious societal challenges.

What, then, can be done to address the challenges of misinformation and disinformation? Potential responses can be roughly divided into three categories. Some responses are individualistic, in the sense that they recommend improved conduct on the part of individuals, perhaps assisted by efforts to improve individual critical thinking skills. Other responses are political. In some jurisdictions, for example, the spread of misinformation and disinformation are criminalized. Finally, some responses center on the platforms by which information is spread. The relevant platforms may include both traditional media outlets and social media platforms. For example, some have proposed that the so-called "Fairness Doctrine" problematically boosted the influence of misleading information by giving fringe views undue prominence in mainstream outlets (McBrayer 2021, p. 31). This paper focuses on an especially contentious response to the challenges of misinformation and disinformation: the removal of content and users from social media platforms. I argue that such measures have significant and underappreciated epistemic downsides. The upshot is not that such measures ought never to be taken, but rather that the effects of such measures are likely to be far less straightforward than is sometimes supposed. Thus, at a minimum, care ought to be taken to implement such measures only in such a way as to minimize their negative consequences.

✉ Keith Raymond Harris
keith.raymond.harris@univie.ac.at

[1] Philosophy Department, University of Vienna, Wien, Austria

Published online: 09 September 2024

The structure of the paper is as follows. Section one provides some brief background on social media misinformation and disinformation and the challenges these pose. Section two discusses the removal of content and users from social media, with a focus on how these measures would seem, at least on their faces, to address the challenges of misinformation and disinformation. Section three raises the concern that, by interfering in the availability of information, policies of user and content removal threaten the perceived value of the information that remains available. Section four provides a theoretical background for this concern, arguing that acts of control over the presentation of evidence function to compromise the force of that evidence. Section five concludes.

2 Bad Information and its Consequences

To begin, it is worth sketching working definitions of misinformation and disinformation. These being technical terms, the ambition of these definitions is not to capture some pretheoretic understanding of their referents. The aim is the more modest one of fixing our targets in the discussion to come. The definitions given are thus stipulative, but aim to capture to a large extent what is typically understood as misinformation and disinformation.

Some philosophers have proposed to understand misinformation as a subset of false claims (De Ridder 2021; Millar 2022). For present purposes, such a definition is too narrow. While some forms of misinformation—for example, fake news articles and bogus scientific findings—may plausibly be understood as consisting in claims, other forms—for example, deepfakes and doctored photos—are not naturally understood in terms of claims. For this reason, I opt for a more inclusive approach to misinformation that is inspired by Don Fallis and Kay Mathiesen's understanding of fake news as "counterfeit news" (2019). On the approach adopted here, fake news is a form of misinformation because it is the "fake" counterpart of a legitimate form of information—in this case news reports generated by truth-conducive processes like fact-checking and editorial review[1]. But, on the present approach, fake news is just one form of misinformation. The category of misinformation includes deepfakes, doctored photos, and fraudulent scientific publications—all of which are "fake" counterparts of more reliable forms of information (cf. Harris 2022).

Disinformation is commonly distinguished from misinformation at least in part by reference to the intentions of some or all of its disseminators. Some have proposed, for example, that disinformation, unlike misinformation, necessarily involves an intention to cause false beliefs (Floridi 2011, p. 260; Jaster and Lanius 2021). Others have suggested that this definition might be too narrow, as some disinformation aims at preventing true beliefs, rather than causing false beliefs (Fallis 2015, p. 420). Finally, it has been suggested that disinformation may not aim at influencing beliefs at all, but may instead target "sub-doxastic associations" (Harris 2023).

For present purposes, we need not take a stand on precisely how broadly to define disinformation. What matters is that, whereas misinformation can be produced and disseminated accidentally, disinformation involves malicious intent on the part of (some of) its promotors. A piece of disinformation might be intrinsically indistinguishable from a piece of misinformation—indeed the same item might be misinformation at one time and disinformation at another—but is distinguished by the intentions of (some of) those who spread it. For example, a misleading scientific publication might be produced with no deceptive intent, but subsequently shared with the intention of deceiving readers into believing a particular false conclusion. Because misinformation and disinformation may be intrinsically indistinguishable, I focus in what follows on misinformation, with the understanding that what is said about the effects of misinformation and the removal of misinformation will, by and large, also apply to disinformation. It is consistent with this basic point that disinformation may be in some sense more concerning than misinformation, insofar as it involves a deliberate and targeted attempt to manipulate an audience.

Misinformation and disinformation are associated with various negative consequences. Some such consequences are dramatic. For example, it is plausible that the storming of the US Capitol in January 2021 was a consequence of misinformation and disinformation suggesting that the 2020 US Presidential election was rigged against Donald Trump. Less dramatically, but perhaps more impactfully, some resistance to the COVID-19 vaccine, and vaccines more generally, is plausibly due to misinformation and disinformation about the safety of vaccines.

Here, I focus mainly on the *epistemic* consequences of misinformation and disinformation. The epistemic consequences are those bearing on the acquisition or retention of epistemic goods. For example, I (2022) have argued that misinformation can interfere with the three conditions often thought to be necessary for knowledge. The truth condition is threatened insofar as misinformation and disinformation promote beliefs in falsehoods, rather than truths. The belief condition is threatened insofar as concerns about the possibility of being misled by misinformation and disinformation lead to hesitation to form beliefs based on even legitimate

[1] Jessica Pepp, Eliot Michaelson, and Rachel Katharine Sterken (2019) defend a related view of fake news, according to which fake news is distinguished by being improperly treated as having been produced through ordinary journalistic practices by those who spread it.

information. Threats to the warrant condition depend on how warrant is analyzed. Misinformation and disinformation may, for some examples, threaten warrant by reducing the amount of information carried by legitimate forms of information (cf. Fallis 2021), by introducing relevant alternatives (Blake-Turner 2020), by reducing the reliability of belief-forming processes (Fallis 2021, p. 625), by making it a matter of luck whether one forms true beliefs (Harris 2021; n. 3; Matthews 2023), and so on.

Knowledge is not the only epistemic good threatened by misinformation and disinformation. For example, Jeroen de Ridder (2021) has argued that misinformation can function to defeat justification for beliefs about dependency relations and, as a consequence, can undermine understanding. Beyond threats to individual knowledge and understanding, misinformation and disinformation threaten *collective epistemic goods*. This is especially clear on reductionist analyses of collective epistemic states. Suppose, for example, that collective knowledge is understood *summatively*, such that a population knows that *p* just in case all or most of its members know that *p*. On such an account of collective knowledge, misinformation and disinformation can interfere with collective knowledge only by interfering with knowledge at the individual level.

If collective epistemic states are not understood in this summative way, misinformation and disinformation can interfere with collective knowledge even independently of their effects on individual knowledge. Consider an example. Some social epistemologists argue that there is a form of collective knowledge that is partly realized in materials beyond the organismic boundaries of any human members of the collective. Bird (2010, 2014), for example, has argued that scientific knowledge is largely realized in material stores of information like journal articles. Similarly, one might argue that there is a form of collective knowledge that is stored in online content, including on social media, independent of being represented in any human individual's mind. On such an account of collective knowledge, the existence of misinformation and disinformation on social media would seem to directly preclude collective knowledge—at least about subjects tainted by high degrees of misinformation and disinformation—even independently of any human person internalizing it.

In this section, I have provided a brief overview of social media misinformation and disinformation and a theoretical account of how these might interfere with individual and collective epistemic goods. The extent to which misinformation and disinformation *actually* interfere with these goods is a matter for empirical study. It is worth noting, in this connection, that a raft of recent work in philosophy and cognitive science suggests that the epistemic ill-effects of misinformation and disinformation are more limited than is commonly supposed. For example, Mercier (2020) has argued that individuals are less susceptible to deception by misinformation than is commonly thought. Additionally, a growing body of research suggests that misinformation is principally consumed and shared by a small fraction of social media users (Grinberg et al. 2019; Osmundsen et al. 2021) and that interactions with notorious promoters of disinformation like Russian trolls are both concentrated among a small number of users and ineffective at changing political behaviors (Eady et al. 2023).

Still, to this point, the effectiveness of social media misinformation and disinformation remains understudied. Even if certain dramatic effects of such misinformation and disinformation are relatively rare, more subtle effects might be relatively common. For example, even if few people are tricked into believing falsehoods by social media misinformation and disinformation, the effects of misinformation and disinformation on the belief and warrant conditions on knowledge might be more dire. Moreover, even if the ill-effects of misinformation and disinformation are concentrated among a small number of users, these effects might be concerning nonetheless. For one thing, even relatively small numbers of misinformed voters might prove decisive in close elections (van der Linden 2023). For another, misinformed beliefs sometimes lead to acts of extremism. Thus, although the ill-effects of misinformation and disinformation warrant more careful empirical study, there is already good reason to consider how these effects might be mitigated.

3 Content Moderation

Given the harms associated with social media misinformation and disinformation, it is no surprise that various proposals have been made as to how these ill-effects might be mitigated. Some such proposals are purely individualistic. For example, some researchers have attempted to create techniques for "inoculating" the public against misinformation (van der Linden 2023). These techniques are intended to function analogously to vaccines, by giving individuals "weakened doses" of misinformation that increase their resistance to subsequent misinformation encountered in the real world. One concern about this and similar measures is that there may well be a *spillover effect* (Van Der Meer et al. 2023), whereby individuals become more resistant not only to misinformation, but to legitimate information as well (Jungherr and Rauchfleisch 2024; Modirrousta-Galian and Higham 2023; Van Duyn and Collier 2019). Another concern is that research on the so-called *truth effect* suggests that repeated exposure to misinformation increases the fluency with which its contents are processed, thereby

promoting credulity toward it (Dechêne et al. 2010; Hassan and Barber 2021). Thus, even without mistaking misinformation for credible information, one might slowly develop credulity toward misinformation. This latter consideration suggests that the best measures for mitigating misinformation may be those that decrease exposure to it.

Reducing exposure to misinformation might be accomplished in various ways. In principle, states might criminalize the distribution of misinformation, thereby deterring its spread. Several considerations, including the potential for abuse and misapplication, as well as legal protections of free speech, militate against such a proposal. However, it is often hoped that social media content moderation—including the removal of misinformative content and de-platforming of misinformation spreaders—can achieve similar results. While concerns about abuse, misapplication, and free speech protections can also be raised against social media content moderation, these critiques have somewhat less force when restrictions on speech are carried out by, and within the relatively narrow context of, privately-owned platforms. Here, I focus specifically on the *epistemic* merits and demerits of content moderation processes.

Often, content moderation is carried out for reasons having little to do with promoting epistemic outcomes. For example, content moderation is used to remove hate speech, calls to violence, and pornography from major social media platforms. Such interventions are mainly aimed at reducing the toxicity of social media platforms, rather than promoting a healthy epistemic environment. In contrast, content moderation that takes the form of labeling and removing misinformation from social media platforms is more plausibly aimed at producing good epistemic outcomes[2]. This is not necessarily to say that platforms, their owners, or their agents are themselves motivated by the aim of improving epistemic environments. However, insofar as social media platforms remove or label misinformation in order to address the concerns of users and advertisers that have epistemic motivations, the resultant content moderation is in some sense motivated by epistemic aims. Although epistemically-aimed content moderation often takes the form of labeling dubious content, or algorithmically deprioritizing it, I focus here on the removal of such content and those who spread it. Still, as I will emphasize below, these alternative forms of content moderation face some challenges very similar to those highlighted here.

One obvious benefit of removing misinformation and its spreaders is that, by doing so, platforms reduce the chances of users being deceived by the content in question. Take, for example, the notorious case of Edgar Maddison Welch, who was arrested after firing shots in a Washington D.C. area pizza restaurant. Shortly before his arrest, Welch was convinced by YouTube videos of the outlandish "Pizzagate" conspiracy theory, according to which Democratic politicians and other elites used the restaurant as a base of operations for a child sex trafficking ring. Although many profess to believe in Pizzagate, or the broader QAnon conspiracy theory into which Pizzagate has been folded, Welch's case is relatively unique in two respects. First, scholars have questioned whether professions to believe in QAnon are sincere, or instead serve to signal political loyalties. The latter conclusion is motivated by the fact that many professed believers seem content to behave in ways that appear misaligned with the seriousness of the allegations at the heart of that conspiracy theory. For example, Hugo Mercier has noted that the Pizzagate allegations led many supposed believers to leave negative online reviews for the pizza place at its heart (Mercier 2020, p. 260). Such petty behaviors seem at odds with the gravity of the Pizzagate claims. In contrast, Welch's actions, although misguided, are aligned with the seriousness of the allegations he professed to believe. In other words, Welch seems to be a clear case of someone actually believing the content of misinformation, rather than using the content of misinformation as a means of signaling. Second, while the causal effects of misinformation are often gradual and diffuse, Welch's case involves what appears to be a fairly straight line from exposure to misinformative YouTube videos to, just days later, action predicated on false beliefs based on these videos. Thus, in Welch's case, it is relatively easy to identify the causes of his false belief.

Given that Welch's false belief seemed to be both genuine and caused by exposure to a relatively identifiable set of misinformation, it is plausible enough to conclude that his false belief could have been prevented by removing the misinformation in question from YouTube. Although the causal connection between misinformation and false belief is typically more difficult to establish, it is plausible enough that the removal of misinformation and its spreaders from social media reduces exposure to misinformation and, in this way, reduces false beliefs based on misinformation. Additionally, de-platforming spreaders of misinformation is likely to reduce false beliefs by both eliminating some individuals' opportunity to spread misinformation and by deterring the posting of misinformation. Although I will not pursue the point at length here, the algorithmic deprioritization of misinformation could similarly be expected to directly reduce exposure to misinformation and to discourage the creation

[2] As an anonymous referee has pointed out, however, that the aims and consequences of removing misinformation cannot be clearly distinguished from the aims and consequences (Harris 2024, p. 84). For one thing, hate speech may include false claims or other content that amounts to misinformation. For another, the removal of hate speech and other toxic content, or the failure to remove such content, may impact the epistemic weight users assign to the content they encounter on social media.

of misinformative content, albeit to a lesser degree than removal.

The effects of removing misinformation and de-platforming its spreaders with respect to false beliefs are not likely to be uniformly positive. Some of those who are de-platformed, or prevented from accessing the sort of (mis) information they seek, may turn to alternative "dark platforms" and other online spaces that are less moderated (Horta Ribeiro et al. 2021). There, they are likely to encounter highly concentrated misinformation. In this way, content moderation may have the effect of radicalizing some users. Thus, for example, empirical research on discourse among members of fringe communities on dark platforms reflects radicalization over time, and this change is especially apparent among relatively recent communities (Schulze et al. 2022). Still given the vast scale of major social media platforms, and the relative rarity of defections to dark platforms, it is reasonable to expect that the removal of misinformation and de-platforming of its spreaders reduces the number of false beliefs formed based on misinformation.

The removal of misinformation and its spreaders can, at least in principle, likewise safeguard the perceived legitimacy of online content. In this way, these forms of content moderation can preserve trust in accurate online content, thereby maintaining the disposition to form beliefs based on this content. To see this, suppose that one is initially concerned about the possible existence of deepfakes on social media platforms. If one becomes confident that such fabricated videos are swiftly removed from social media, or prevented from being posted in the first place, one will maintain one's trust in the authenticity of video content. More generally, efforts to purge social media of misinformation can, at least in principle, help to maintain credulity toward content that remains available.

There are several complications for this story about the positive effects of removing misinformation from social media. To maintain confidence in online content, it is not enough to remove misinformative content. At a minimum, users must trust that misinformation is effectively removed. If users suspect that efforts to remove misinformation fail to catch a significant proportion of misinformation, their distrust of online content will not be significantly alleviated. Adding to this concern are recent empirical results suggesting that efforts to warn the public about misinformation tend to reduce trust in even accurate information (Van Duyn and Collier 2019). Thus, even if efforts to remove misinformation are in fact effective, these same efforts may make the existence of misinformation more salient, thereby reducing public trust in even accurate online content (cf. Lecheler and Egelhofer 2022, p. 82).

Realistically, efforts to remove all misinformation from social media platforms are not likely to catch all extant misinformation. This leads to a further concern. If users *do* trust content moderation processes to remove misinformation, but some misinformation is overlooked, any remaining misinformation is likely to receive a boost in perceived credibility. This phenomenon may be understood as an instance of the *implied truth effect*. This effect has previously been observed as a consequence of labels for misinformative content. Empirical researchers have observed that, when some but not all misinformation is labeled as such, non-labeled misinformation is perceived as credible, relative to a situation in which no misinformation is labeled (Pennycook et al. 2020). Similarly, if some but not all misinformation is removed, and users know about efforts to remove misinformation, they can be expected to ascribe undue credibility to unlabeled misinformation. Thus, while removing misinformation plausibly preserves the credibility of accurate content, insufficiently thorough processes for removing misinformation can be expected to also boost the perceived credibility of misinformation.

These concerns ought not be taken too far. Even if processes for removing misinformation could not be expected to remove *all* misinformation, such processes could be expected to remove major strains of misinformation or misinformation posted by especially influential users. Thus, at least in principle, policies of removing misinformation might help users to retain their trust in content that is heavily circulated, especially by influential users over a long period of time. More concretely, once an item of misinformative content is identified, automated techniques can be used to efficiently remove instances or variants of that content. Thus, policies for removing misinformation might help assure users that a particular highly-circulated video is authentic and not, for example, a deepfake.

Removing misinformation from social media also holds some promise for preserving the ability to obtain warrant from online content. Consider, for example, the view according to which warrant for a given belief requires that the process by which the belief is formed is reliable. In an environment that includes a good deal of misinformation, certain belief-forming processes—for example those that involve reading news reports—will be relatively unreliable. If fake news reports are effectively removed from the epistemic environment, however, the reliability of that belief-forming process will be maintained. By a similar token, the removal of misinformation from social media platforms helps to ensure that it is not simply a matter of luck when one forms true beliefs based on social media content. Similarly, the removal of misinformation, at least if carried out comprehensively, would remove relevant alternatives according to which various online contents are fakes. Consider, finally, the effect of removing misinformation on the informational value of online contents. Fallis's contention

that deepfakes reduce the amount of information conveyed by video footage is based on the following account of information, described by Skyrms (2010):

> [A] signal R carries information about a state of affairs S whenever it distinguishes between the state of affairs where S is true and the state where S is false. That is, R carries the information that S when the likelihood of R being sent when S is true is greater than the likelihood of R being sent when S is false. (Fallis 2021, p. 629)

If fakes abound, then the probability of there being a signal of some occurrence despite that occurrence not being actual—say, a deepfake depicting an event that did not exist—is relatively high. It is for this reason that deepfakes reduce the informational content of video footage. Thus, removing deepfakes is a way of preserving the informational content of video footage. By a similar token, removing various sorts of fakes from an epistemic environment—or preventing their introduction to the environment in the first place—serves to maintain the informational content of the relevant authentic counterparts, including videos, photos, news reports, and so on.

Content moderation can also plausibly protect epistemic goods beyond individual knowledge and its components. Suppose, following de Ridder (De Ridder 2021), that misinformation threatens understanding by interfering with justification for beliefs about dependency relations. Then, insofar as content moderation safeguards justification, it likewise functions to safeguard understanding. Depending on how collective epistemic states are understood, the protective effects of content moderation with respect to such states may likewise be derivative of protective effects with respect to other epistemic goods. Suppose, for example, that collective knowledge is understood summatively. Then, to the extent that content moderation protects individual knowledge from the threats of misinformation, it likewise protects collective knowledge. What about proposals on which collective knowledge may be partially realized in non-human material stores of information? On such proposals, misinformation and disinformation that is realized in non-human stores of information can plausibly prevent collective knowledge by, for example, partly constituting false collective beliefs or simply acting as noise that obscures accurate social media content. Thus, on such proposals, the purging of misinformation and disinformation from social media can plausibly serve to preserve the collective knowledge that is partly constituted by social media contents.

4 Some Psychological Consequences of the Removal of Misinformation

Despite the various epistemic benefits of removing misinformation from social media, as described in the preceding section, I now argue that there are serious and underappreciated epistemic risks for the removal of misinformation from social media. I begin, in this section, by arguing that removing misinformation is likely to cause or exacerbate skepticism about the evidential value of certain kinds of information encountered on social media. In other words, I argue that removing misinformation is likely to have the effect, among some social media users, of causing or exacerbating distrust of certain kinds of information. It is worth emphasizing that the effect I describe in this section is a *psychological* consequence of the removal of misinformation from social media. In Sect. 4, I argue that this psychological effect is, at least in part, *rational*.

To start, consider two kinds of information that are commonly obtained by social media users. Some information is obtained from social media content. One might learn something from (or be misinformed by) the content of a news article, a photo, a video, and so on. For example, one might learn, by watching a video posted to social media, that a given politician made a gaffe during a campaign appearance. Another sort of information is obtained from the social context surrounding social media content. For example, by observing that a given scientific article has been extensively shared with positive reactions by members of the relevant scientific discipline, and has rarely or never been disparaged, one might learn that the claims made in the article are considered plausible by members of the discipline.

The distinction here is between the information contained in social media content itself and the information conveyed by the social context of that content. Typical beliefs based on the former will concern whatever the social media content is about. For example, if the social media content in question is a news article about the economy, resultant beliefs will concern the economy. In contrast, typical beliefs based on the latter will concern the attitudes of social media users toward the content in question. For example, if one observes many working particle physicists reacting negatively to a new article on the possible discovery of a new elementary particle, one might conclude that there is no consensus in particle physics supporting the claims made in the article. This distinction between the beliefs formed based on social media content and its surrounding social context is only rough. For example, when one reads an article posted to social media by its author, one is likely to form beliefs not only about the contents of the article, but also about the attitudes of its author. For another, the social context surrounding a piece of social media content may determine whether

or not one forms a belief in alignment with that content. To return to the example above, the apparent disapproval of other particle physicists might lead one to avoid forming the belief that a novel elementary particle has been discovered.

In Sect. 2, I argued in effect that content moderation can—complications aside—function to protect the integrity of beliefs based on information of this first kind. According to that line of argument, removing misinformation from social media can be expected to promote the formation of beliefs based on accurate content and, in this way, can promote the acquisition of knowledge and other epistemic goods. On the face of things, it might appear that the removal of misinformation from social media can likewise protect the integrity of beliefs based on information of the second kind. Unless misinformation is removed from social media, misleading social contextual information might lead to false beliefs about both users' attitudes and the accuracy of social media content. Suppose, for example, that a misleading scientific article appears on social media and that a small but highly vocal minority of scientists—wrongly, but not necessarily maliciously—express their support for its claims. In this case, laypersons on social media might be easily misled about both what the overall scientific attitude toward the article is and whether the conclusions reached in the article are correct. Thus, one might think, the removal of misinformation from social media would serve to remove confusion about both the accuracy of content and the attitudes of others.

However, this story is overly simple. The removal of misinformation from social media does not simply function to purge the epistemic environment of potentially misleading content and social contextual information. Rather, removing misinformation threatens to create the perception that what is allowed to appear on social media is subject to manipulation and thus that information of the second kind—namely information about others' attitudes toward social media content—cannot be taken at face value. Consider an example inspired by the perception of the effects of content moderation during the COVID-19 pandemic. Suppose that misinformation falsely alleging the mildness of a virus, the dangers of mainstream treatments, and the effectiveness of alternative treatments is systematically removed from social media and its promoters, including individuals with relevant credentials, are de-platformed. On the one hand, this is likely to have many of the epistemic benefits I associated with content moderation in Sect. 2. On the other hand, at least for some social media users, one effect of such policies is likely to be the perception that social contextual information that remains on social media is not representative of overall opinion[3]. For example, some social media users may be suspicious that what appears to be the consensus position among those with relevant credentials and the public more widely is an artificial and illegitimate consensus.

This preceding line of argument is not merely speculative. Harambam (2023) presents evidence of conspiratorial suspicions about scientific consensuses concerning COVID-19 that are predicated on, among other things, concerns about the suppression of dissent on social media. Notably, Harambam's evidence comes from individuals previously labeled as conspiracy theorists. Thus, one might argue that such evidence serves only to show that the suppression of dissent only drives suspicions about the legitimacy of consensuses among those who are already inclined to conspiracy theorizing. One might argue, further, that creating suspicions among such audiences is no objection to the policy of suppressing misinformation, as virtually any occurrence can be interpreted by the conspiratorially-minded as part of a conspiracy (cf. Keeley 1999).

There are at least two problems with this dismissive line of objection. The first is that suspicions about the legitimacy of consensuses influenced by the suppression of dissent are hardly unique to fringe conspiracy theorists. Within the philosophy of science, for example, it is has often been suggested that the social evidential weight of consensuses is contingent on the possibility of dissent (De Melo-Martín and Intemann 2018; Intemann 2017; Oreskes 2021, p. 32). More directly, there is reason to think that concerns about the effect of content moderation on the epistemic value of consensus are, to some degree, reasonable. I make the case for this point in the following section. If this argument is sound, there is further reason to deny that doubts about the value of consensus following the suppression of dissent are restricted to individuals who start with conspiratorial suspicions.

5 Some Epistemic Consequences of the Removal of Misinformation

In this section, I make the case that the removal of misinformation can negatively impact the evidential value of information that remains available on social media. This argument may be surprising in light of the argument developed in Sect. 2, according to which the removal of misinformation from social media can function to preserve the ability of accurate content—in the form of photos, videos, news reports, and so on—to carry information. As I will argue in what follows, the argument given in Sect. 2 relies on an oversimplification concerning the factors bearing

[3] Notably, even if misinformation is not outright removed from social media, subtler policies of algorithmic suppression might likewise give the impression that the social contextual information accessible on social media underestimates the credibility of the affected misinformation.

on the weight of evidence. A more sophisticated approach to evidence will make clear a potential drawback of the removal of misinformation.

To start, consider the evidence that might be available to a social media user in a particular instance. Suppose that there is a strong consensus among relevant epistemic authorities—that is, those holding relevant professional positions and credentials—that a given substance, X, offers no benefits for preventing contraction, serious illness from, or transmission of a certain novel virus, and has some serious side effects. Suppose that this is in fact true. Suppose, however, that a small minority of relevant epistemic authorities insist that X effectively reduces transmission of the virus, the severity of its effects, and causes no significant side effects, and spread misinformation to this effect. Laypersons, in virtue of lacking relevant statistical and domain expertise, may be incapable of assessing the first-order evidence bearing on the effectiveness and safety of X. Nonetheless, they may use the judgments of relevant epistemic authorities as indicators of the truth about X.

What should a layperson believe in such a case? Supposing that the layperson has no antecedent reason to think that some of the epistemic authorities in question are more reliable than others, it would be reasonable for the layperson to conclude that those epistemic authorities in the majority are correct, and thus that X does not effectively prevent contraction, serious illness, or transmission of the virus and has some serious side effects. One might argue for this conclusion by reference to formal results indicating the reliability of majority judgments among suitably large populations. However, this conclusion cannot rely straightforwardly on Condorcet Jury Theorem-style reasoning, as it is unlikely that the independence assumption that is necessary for application of that theorem is likely to hold in the relevant case (Goldman 2001, pp. 99–104). Especially among epistemic authorities in a particular domain, it is likely that individuals exhibit mutual influence on one another's judgments. Still, even if the independence condition is not met, a relatively large body of epistemic authorities—those representing the majority—is likely to be more reliable than the minority. Even if individuals influence one another, they are not likely to be uncritical adopters of the beliefs of their colleagues (Lackey 2021, pp. 209–214). Moreover, each may consider possibilities for error not considered by the others. In short, there is reason to expect the majority position to have withstood greater scrutiny than the minority position. Thus, other things being equal, a given layperson has more reason to accept the majority view than the minority view.

It might be thought that, whatever support the judgments of those in majority lend to the proposition that X is ineffective and potentially dangerous, this support is tempered by the competing judgments of those in the minority. Typically, dissent from a widely-accepted view seems to offer some reason to doubt that view or, more modestly, to reduce one's degree of belief in its truth. This, one might think, is especially true in cases like the one described here, in which dissent is offered by individuals who laypersons have antecedent reason to believe are as reliable as the epistemic authorities in the majority. Just as one ought to reduce one's confidence in one's own judgments when these are met by disagreement from one's peers, disagreement among epistemic authorities seems to be a reason to limit one's belief in the claims of epistemic authorities—even those representing the majority.

Given the negative epistemic effects of dissent on the case for believing the majority view, one might think that there is good reason for third parties to suppress the visibility of dissent. In effect, one might think, suppressing dissent in this way would serve to limit the availability of misleading social evidence, thereby strengthening the epistemic position of laypersons. For example, social media platforms might limit or eliminate the accessibility of expert judgments supporting the claim that X is safe and effective introduced above. In the previous section, we saw reason to think that such a policy might backfire, by making social media users suspicious of the information that remains available on social media platforms. In particular, social media users might begin to suspect that apparent consensus supporting the ineffectiveness and risks of X is illegitimate. Such *psychological* consequences, I now argue, may reflect a legitimate *epistemic* ill-effect of suppressing misinformation.

I have stipulated above that, in the case introduced here, it is a fact that X is ineffective and potentially dangerous. Thus, one might think—in line with an argument presented in Sect. 2—that removing dissent from this by epistemic authorities would amount to removing pollution from the epistemic environment, thereby increasing the informational value of the claims and evidence of epistemic authorities representing the majority position. But this is too quick. Allowing the claims of epistemic authorities—even false claims—to be suppressed, introduces several epistemic challenges.

To see this, consider first the relationship between a layperson that comes to believe that X is ineffective and unsafe based on the testimony of a given relevant epistemic authority. This relationship is characterized by a paradigmatic form of epistemic dependence (Broncano-Berrocal and Vega-Encabo 2020; Hardwig 1985). Not only is the layperson's belief *caused* by the testimony of the epistemic authority, the epistemic properties of that layperson's belief—in particular the warrant for it—depend on features of the epistemic authority. For example, the degree to which the layperson's belief is warranted plausibly depends on the competence and sincerity of the epistemic authority. Now

suppose that, rather than trusting fully in any particular epistemic authority, the layperson attends to a large profile of judgments from relevant epistemic authorities—most of whom contend that X is ineffective and unsafe—and, as a consequence, comes to form the corresponding belief. In this case, the layperson is not highly dependent on any particular epistemic authority, but remains dependent on such authorities as a group. Such a case involves a kind of *diffuse* epistemic dependence (cf. Goldberg 2011). Forms of epistemic dependence, whether one-on-one or diffuse, involve a kind of vulnerability on the part of the dependent. Properties of the epistemic authority or authorities, including proneness to mistakes or dishonesty, might epistemically harm the dependent. However, supposing that positions of epistemic authority are typically reserved for those who are especially competent, suffering such vulnerability is typically a small price to pay for the attendant epistemic benefits of epistemic dependence of others.

Consider, however, one further wrinkle on the case. Suppose that third-parties—in this case, content moderators—regularly intervene in the content available to laypersons. We might suppose, for example, that content moderators have a policy of removing misinformation. In this case, laypersons are epistemically dependent not merely on producers of content—including epistemic authorities—but also on those who moderate content. There are various ways in which dependence on content moderators might leave laypersons worse off. For example, if content moderators are ill-motivated—aiming to suppress the truth and promote falsehoods, for example—then they might cherry pick for promotion content from epistemic authorities whose outputs serve these ends, while removing conflicting claims. Forming beliefs based on content available on social media would, in this case, be an unreliable method of belief formation. While some social media users certainly suspect that content moderation sometimes works in this way, such conspiratorial suspicions need not be accurate in order for social media users to be made epistemically worse off by social media content moderation. Content moderators might, through good-faith error, wrongly remove accurate content.

Importantly, content moderators need not *actually* remove accurate content in order for social media users to be made worse off by the existence of programs of content moderation. The mere realistic possibility that they might do so introduces possibilities for error that can interfere with the warrant condition on knowledge. To see this, consider some approaches to warrant discussed above. According to one such approach, knowledge is inconsistent with the existence uneliminated relevant alternatives. Insofar as there exist programs for removing content from social media, a relevant alternative for any particular belief formed based on what appears to be the consensus of epistemic authorities encountered online is that such a consensus was manufactured through the suppression of dissent by content moderators. Now consider the view according to which warrant requires that beliefs be formed through reliable processes. Ordinarily, forming beliefs based on the apparent consensus judgments of epistemic authorities is a highly reliable process. Consider, for example, scientific beliefs formed in this way. Because the relevant epistemic authorities—scientists—are generally highly competent, and because their outputs tend to be subjected to truth-conducive processes including peer review, believing based on the apparent consensus of relevant scientists is typically a highly reliable process. In contrast, if the appearance of consensus can be manufactured by a relatively small number of content moderators who lack expertise in the domain and whose work is not subject to peer review, then believing based on apparent consensus will be a comparatively unreliable process. This latter point can also be put in terms of the information conveyed by consensus. Ordinarily, apparent consensus among relevant epistemic authorities is a strong signal of the truth, insofar as such a consensus is unlikely to emerge in favor of a falsehood. However, content moderation makes it easier to generate a misleading apparent consensus and, for this reason, reduces the informational value of apparent consensus.

Often, epistemologists invoke the concept of epistemic dependence to describe situations involving two roles—deliverers and recipients of information. What I have highlighted here is that epistemic dependence can also involve three roles—deliverers, recipients, and third-party controllers of information. In some cases, third-party controllers of information can play a vital role in promoting epistemic goods for the recipients of information. For example, Goldberg (2007; Chap. 8). describes a case in which a parent limits the sort of testimony that is able to reach her child's ears according to its accuracy, thereby safeguarding the reliability of the child's testimonial beliefs. In effect, such a parent controls the social evidence available to the child. In the case described, such *social evidence tampering* plausibly epistemically benefits the child. Likewise, content moderation construed as social evidence tampering undoubtedly has some benefits, some of which are described in Sect. 2. However, as I have argued in this section, social evidence tampering in the social media context can also be costly with respect to the warrant condition on knowledge.

6 Concluding Remarks

In this paper, I have provided an overview of some benefits and costs of social media content moderation with respect to the aim of promoting knowledge. While the benefits of

content moderation are straightforward, the costs are more subtle. The removal of misinformation from social media can help to remove confusion about what is true, thus promoting the acquisition of warranted true belief. However, such removals require the exercise of power over the evidence available to others. Even the specter of such power can be enough to encourage suspicions about its exercise and can thus discourage credulity toward even accurate information encountered on social media. Such power can also compromise the value of social evidence, thereby interfering with the warrant condition on knowledge.

While I have presented some benefits and costs of content moderation, I have not attempted to weigh these against one another, and thus I offer no judgments on whether the removal of misinformation from social media is justified. A well-reasoned judgment on this point must be sensitive to epistemological considerations, as well as ethical and practical ones. Moreover, any such judgment should account for at least three further facts. First, even if content moderation is not used against misinformation, it is likely to be used against hate speech, calls to violence, and so on. Exercises of content moderation in these contexts are likely to provoke suspicions on the part of some social media users, and thus the encouragement of such suspicions is not a cost associable only with the removal of misinformation. Second, removal is not the only way of addressing the challenge of misinformation. Arguably, other strategies like labeling suspected misinformation are better alternatives to removal. Third, even if content moderators do not exercise control over the epistemic environment on social media platforms, others will. In fact, part of what is distinctive about the epistemic environments of social media platforms, as opposed to more traditional forms of media, is their susceptibility to control—in the form of the addition and promotion of content—by ordinary individuals. While ordinary individuals have a very limited ability to control the content of social media, relatively novel technologies like generative AI and bot networks enhance the ability of motivated actors to exercise a relatively high degree of control over the evidence available on social media. In light of these and other considerations, the present essay serves only to highlight some issues that ought to be taken into account in judgments about the case for and against exercises of content moderation.

Funding This paper was partly written in the context of the "INTERACT!" project, which has received/is receiving funding from the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. This research was also funded in part by the Austrian Science Fund (FWF) [https://doi.org/10.55776/COE3]. For open access purposes, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. The sole responsibility for the content of this publication lies with the author.

Open access funding provided by University of Vienna.

Data availability Not applicable.

Declarations

Competing Interests The author has no competing interests to report.

References

Bird A (2010) Social Knowing: the Social sense of scientific knowledge. Philosophical Perspect 24:23–56

Bird A (2014) When is there a group that knows? In: Lackey J (ed) Essays in collective epistemology. Oxford University Press, pp 42–63. https://doi.org/10.1093/acprof:oso/9780199665792.003.0003

Blake-Turner C (2020) Fake news, relevant alternatives, and the degradation of our epistemic environment. Inquiry 1–21. https://doi.org/10.1080/0020174X.2020.1725623

Broncano-Berrocal F, Vega-Encabo J (2020) A taxonomy of types of epistemic dependence: introduction to the Synthese special issue on epistemic dependence. Synthese 197(7):2745–2763. https://doi.org/10.1007/s11229-019-02233-6

De Melo-Martín I, Intemann K (2018) *The Fight Against Doubt* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190869229.001.0001

De Ridder J (2021) What's so bad about misinformation? Inquiry 1–23. https://doi.org/10.1080/0020174X.2021.2002187

Dechêne A, Stahl C, Hansen J, Wänke M (2010) The Truth about the truth: a Meta-Analytic Review of the Truth Effect. Personality Social Psychol Rev 14(2):238–257. https://doi.org/10.1177/1088868309352251

Eady G, Paskhalis T, Zilinsky J, Bonneau R, Nagler J, Tucker JA (2023) Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nat Commun 14(1):62. https://doi.org/10.1038/s41467-022-35576-9

Fallis D (2015) What is Disinformation? Libr Trends 63(3):401–426

Fallis D (2021) The epistemic threat of Deepfakes. Philos Technol 34(4):623–643. https://doi.org/10.1007/s13347-020-00419-2

Fallis D, Mathiesen K (2019) Fake news is counterfeit news. Inquiry 1–20. https://doi.org/10.1080/0020174X.2019.1688179

Floridi L (2011) The philosophy of information. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001

Goldberg SC (2007) Anti-individualism: mind and Language, Knowledge and Justification, 1st edn. Cambridge University Press. https://doi.org/10.1017/CBO9780511487521

Goldberg S (2011) The Division of Epistemic Labor. Episteme 8(1):112–125. https://doi.org/10.3366/epi.2011.0010

Goldman AI (2001) Experts: which ones should you trust? Philos Phenomenol Res 63(1):85–110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706

Harambam J (2023) Distrusting Consensus: how a Uniform Corona Pandemic Narrative fostered suspicion and conspiracy theories. J Digit Social Res 5(3):109–139. https://doi.org/10.33621/jdsr.v5i3.143

Hardwig J (1985) Epistemic Dependence. J Philos 82(7):335. https://doi.org/10.2307/2026523

Harris KR (2021) Video on demand: what deepfakes do and how they harm. Synthese 199(5–6):13373–13391. https://doi.org/10.1007/s11229-021-03379-y

Harris KR (2022) Real fakes: the Epistemology of Online Misinformation. Philos Technol 35(3):83. https://doi.org/10.1007/s13347-022-00581-9

Harris KR (2023) Beyond Belief: On Disinformation and Manipulation. *Erkenntnis*. https://doi.org/10.1007/s10670-023-00710-6

Harris KR (2024) Misinformation, Content Moderation, and Epistemology: Protecting Knowledge (1st ed.). Routledge. https://doi.org/10.4324/9781032636900

Hassan A, Barber SJ (2021) The effects of repetition frequency on the illusory truth effect. Cogn Research: Principles Implications 6(1):38. https://doi.org/10.1186/s41235-021-00301-5

Horta Ribeiro M, Jhaver S, Zannettou S, Blackburn J, Stringhini G, De Cristofaro E, West R (2021) Do platform migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. Proc ACM Hum Comput Interact 5CSCW2:1–24. https://doi.org/10.1145/3476057

Intemann K (2017) Who needs Consensus anyway? Addressing Manufactured Doubt and Increasing Public Trust in Climate Science. Public Affairs Q 31(3):189–208

Jaster R, Lanius D (2021) Speaking of fake news: definitions and dimensions. In: Bernecker S, Flowerree AK, Grundmann T (eds) The epistemology of fake news. Oxford University Press, pp 19–45

Jungherr A, Rauchfleisch A (2024) Negative downstream effects of Alarmist Disinformation discourse: evidence from the United States. Polit Behav. https://doi.org/10.1007/s11109-024-09911-3

Keeley BL (1999) Of conspiracy theories. J Philos 96(3):109. https://doi.org/10.2307/2564659

Lackey J (2021) Echo chambers, fake news, and Social Epistemology. In: Bernecker S, Flowerree AK, Grundmann T (eds) The epistemology of fake news. Oxford University Press, pp 208–227

Lecheler S, Egelhofer JL (2022) Disinformation, misinformation, and fake news: understanding the Supply side. Knowledge Resistance in High-Choice Information environments. Routledge, pp 69–87

Matthews T (2023) Deepfakes, fake barns, and knowledge from videos. Synthese 201(2):41. https://doi.org/10.1007/s11229-022-04033-x

McBrayer JP (2021) Beyond fake news: finding the truth in a world of misinformation. Routledge, Taylor & Francis Group

Mercier H (2020) Not born yesterday: the Science of who we trust and what we believe. Princeton University Press. https://doi.org/10.1515/9780691198842

Millar B (2022) Epistemic obligations and free speech. Analytic Philos phib12279. https://doi.org/10.1111/phib.12279

Modirrousta-Galian A, Higham PA (2023) Gamified inoculation interventions do not improve discrimination between true and fake news: reanalyzing existing research with receiver operating characteristic analysis. J Exp Psychol Gen. https://doi.org/10.1037/xge0001395

Oreskes N (2021) Why trust science? Princeton University Press

Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. Am Polit Sci Rev 115(3):999–1015. https://doi.org/10.1017/S0003055421000290

Pennycook G, Bear A, Collins ET, Rand DG (2020) The Implied Truth Effect: attaching warnings to a subset of fake News headlines increases Perceived Accuracy of headlines without warnings. Manage Sci 66(11):4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pepp J, Michaelson E, Sterken R (2019) What's New about fake news? J Ethics Social Philos 16(2). https://doi.org/10.26556/jesp.v16i2.629

Schulze H, Hohner J, Greipl S, Girgnhuber M, Desta I, Rieger D (2022) Far-right conspiracy groups on fringe platforms: a longitudinal analysis of radicalization dynamics on Telegram. Convergence: Int J Res into New Media Technol 28(4):1103–1126. https://doi.org/10.1177/13548565221104977

Skyrms B (2010) *Signals: Evolution, Learning, and Information* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780199580828.001.0001

van der Linden S (2023) Foolproof: why misinformation infects our minds and how to build immunity. W.W. Norton & Company

Van Der Meer TGLA, Hameleers M, Ohme J (2023) Can fighting Misinformation have a negative spillover effect? How warnings for the threat of Misinformation can decrease General News credibility. Journalism Stud 24(6):803–823. https://doi.org/10.1080/1461670X.2023.2187652

Van Duyn E, Collier J (2019) Priming and fake news: the effects of Elite discourse on evaluations of News Media. Mass Communication Soc 22(1):29–48. https://doi.org/10.1080/15205436.2018.1511807