Erkenntnis
https://doi.org/10.1007/s10670-023-00710-6

ORIGINAL RESEARCH

Beyond Belief: On Disinformation and Manipulation

Keith Raymond Harris[1]

Received: 27 September 2022 / Accepted: 3 June 2023

Abstract

Existing analyses of disinformation tend to embrace the view that disinformation is intended or otherwise functions to mislead its audience, that is, to produce false beliefs. I argue that this view is doubly mistaken. First, while paradigmatic disinformation campaigns aim to produce false beliefs in an audience, disinformation may in some cases be intended only to prevent its audience from forming true beliefs. Second, purveyors of disinformation need not intend to have any effect at all on their audience's beliefs, aiming instead to manipulate an audience's behavior through alteration of sub-doxastic states. Ultimately, I argue that attention to such non-paradigmatic forms of disinformation is essential to understanding the threat disinformation poses and why this threat is so difficult to counter.

1 Introduction

Recent political developments in the western world have increased public consciousness of disinformation. Together with related phenomena, especially conspiracy theories, disinformation has been implicated in the 2016 Brexit vote (Chivvis, 2016: 5; Cooper, 2021), the 2016 election of Donald Trump to US president (Benkler et al., 2018; Mueller, 2019), resistance to various public health measures in the face of the COVID-19 pandemic (Loomba et al., 2021; Pereira et al. 2020; Tagliabue et al., 2020), and so on. While disinformation is not a novel phenomenon, and many regions of the world have suffered due to disinformation, such recent developments have, for many westerners, brought the destructive potential of disinformation to the

✉ Keith Raymond Harris
keithraymondharris@gmail.com

[1] Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany

Published online: 30 June 2023

fore. Moreover, while disinformation campaigns are nothing new, novel technologies—especially social media—have changed how such campaigns are conducted[1].

The heightened prominence of disinformation in public consciousness suggests that there is value in clarifying what precisely disinformation is and why it poses a threat. Existing analyses of the concept tend toward the view that disinformation is intended or otherwise functions to mislead its audience, that is, to produce false beliefs. I argue in what follows that this view is doubly mistaken. First, while paradigmatic disinformation campaigns aim to produce false beliefs in an audience, disinformation may in some cases be intended only to prevent its audience from forming true beliefs. Moreover, purveyors of disinformation need not intend to influence an audience's beliefs at all, and may instead aim to manipulate an audience's behavior through alteration of sub-doxastic states. Ultimately, I argue that attention to such non-paradigmatic forms of disinformation is essential to understanding the threat disinformation poses and why this threat is so difficult to counter.

2 False Belief Accounts of Disinformation

In this section, I survey several existing accounts of disinformation. The purpose is not to analyze these accounts at length, but to highlight the accounts' shared commitment to the claim that disinformation functions to produce false beliefs. This commitment is constitutive of what I call *false belief accounts* of disinformation. Ultimately, I will argue that false belief accounts are untenable.

Consider, first, how Luciano Floridi distinguishes between misinformation and disinformation:

> [M]isinformation is 'well-formed and meaningful data (i.e. semantic content) that is false'. 'Disinformation' is simply misinformation purposefully conveyed to mislead the receiver into believing that it is information. (2011: 260)

Here, Floridi understands disinformation as the subcategory of misinformation that is intentionally deceptive. Elsewhere, Floridi suggests it is acceptable to use "disinformation" to refer to "false semantic information…that is disseminated in order to mislead its receiver" (2012: 306). Given Floridi's understanding of information as factive, the former gloss is perhaps preferable[2]. However, the issue need not concern us here. For present purposes, the crucial feature of Floridi's approach is that it defines disinformation as intentionally misleading.

It is worth clarifying what it means to mislead. In the quoted passage, the relevant form of misleading is making the receiver believe, falsely, that the misinformation conveyed is information. This aligns with other contemporary accounts of what it is to mislead. Jennifer Saul, for instance, writes that "for a misleading to occur, the audience must end up with a false belief" (2012: ch. 1, fn. 5). Misleading is thus a

[1] Rini (2019), for instance, has argued that novel technologies function to co-opt ordinary citizens into participation in campaigns of disinformation.

[2] Thanks to an anonymous referee for drawing attention to this point.

success term, such that success consists in leaving the target with a false belief[3]. Floridi's account is thus what I call a false belief account of disinformation.

The view that disinformation aims at causing false beliefs is widely shared among philosophers and other academics. Consider just a few examples:

> Disinformation is fake or inaccurate information that is intentionally spread to mislead and/or deceive. (Shu et al. 2020: 2)
> Disinformation: fabricated and factually incorrect information spread with an intention to deceive the audience. (Glenski, Volkova, & Kumar 2020: 43)
> [D]isinformation is false information spread deliberately to deceive. (Jaster & Lanius 2021: 35)

It will be noted that the examples cited here are only false belief accounts if it is assumed that aiming at deception involves aiming to cause false beliefs[4]. This assumption is plausible, given that both philosophical accounts and dictionary definitions tend to tie deception to causing false beliefs (Fallis, 2010; Mahon, 2007, 2015). Still, as I will discuss in Sect. 3, deception is sometimes understood in broader terms. Thus, because the authors whose views are presented here do not explicitly define deception, it is possible that some may endorse a relatively broad view of disinformation along the lines of the one argued for in Sect. 3.

In some places, Don Fallis (2009a; 2014) defends an account of disinformation that closely resembles Floridi's. He writes for example that:

> [D]isinformation is information that is intentionally misleading. That is, it is information that—just as the source of the information intended—is likely to cause people to hold false beliefs. (Fallis 2014: 137)

Elsewhere, however, Fallis offers a relatively broad account of disinformation. According to Fallis, "disinformation is misleading information that has the *function* of misleading someone" (2015: 413). Fallis is explicit about what misleading involves:

> The second important feature of disinformation is that it is a type of *misleading* information; that is, it is information that is *likely* to create *false beliefs*. (2015: 406)

For Fallis, to say that disinformation has the function of misleading is to say that it is non-accidentally misleading (2015: 413). Fallis offers two potential ways in which disinformation can be non-accidentally misleading. Information may be non-

[3] I take it that one might in principle cause false beliefs in a target without misleading that target. The advanced neurosurgeon of the philosophical imagination might cause false beliefs in a patient through physical manipulation of the patient's brain, without thereby misleading the patient. Mahon (2007) and Fallis (2010) make similar points about deceiving.

[4] Similarly, Søe (2021) defines disinformation as intentionally misleading non-natural information. So long as misleading a target is understood to involve causing to target to form false beliefs, Søe's account is likewise a false belief account of disinformation.

accidentally misleading because it is intended to be misleading. Information that is non-intentionally misleading in this way largely conforms to Floridi's account of disinformation. However, according to Fallis, information may also be non-accidentally misleading "because the source systematically benefits from their being misleading" (2015: 413). Disinformation is thus, on Fallis's approach, a relatively encompassing concept.

It is worth asking whether there is sufficient cause to accept Fallis's somewhat expanded conception of disinformation. To do so, let us consider his examples. Fallis aims to include the deceptive signals of non-human animals within the category of disinformation (2015: 412–413). However, Skyrms (2010)— whose work concerning animal signals Fallis cites—does not refer to such signals as disinformation. More generally, such signals are not typically regarded as forms of disinformation. Thus, without an argument for regarding such signals in this way, the example does little to motivate an expanded conception of disinformation. Fallis offers a second, more compelling example, concerning those people who credulously disseminate false conspiracy theories (2015: 411–412). Given their credulity, such persons do not intentionally cause false beliefs. However, because widespread interest in conspiracy theories can drive attention to such persons, therefore encouraging the further spread of conspiracy theories, the spread of false conspiracy theories can be said to be non-accidentally misleading. Unlike deceptive animal signals, conspiracy theories are often regarded as a form of disinformation. However, one plausible explanation for this is that conspiracy theories are often intended to manipulate the audience for political or other purposes (Cassam, 2019; Harris, 2022). If there is no such intention, it is not clear that conspiracy theories constitute disinformation, as opposed to *misinformation*. In the absence of an argument, Fallis's example thus does not sufficiently motivate an expanded conception of disinformation. I will consequently focus below on arguing against the view that disinformation is intentionally deceptive. Still, the arguments to follow will show that even Fallis's expanded conception of disinformation is too narrow.

Before concluding this section, it is worth considering some recent work on the definition of *fake news*, a concept closely related to disinformation. Some scholars propose a definition that parallels false belief accounts of disinformation. For example, McIntyre (2018), and Rini (2017) take fake news to involve an intention to deceive at least some of its audience. So construed, fake news bears a close similarity to disinformation, as the authors discussed in this section understand it. Paralleling how I will argue that the false belief accounts of disinformation are too narrow, some commentators contend that fake news does not require a deceptive intention. It is therefore worth considering these arguments and how they bear on the present issue.

Romy Jaster and David Lanius (2021) argue that, while some fake news is distributed with an intention to deceive, other fake news is distributed without regard for the truth. This latter form of fake news is akin to *bullshit* (Frankfurt, 2005), rather than lies. A more expansive definition of fake news along these lines is sometimes motivated by the plausible claim that fake news can be intended to generate clicks, thereby producing profit, rather than intended to further some ideological aim[5] (Grundmann, 2020; Jaster & Lanius, 2021: 22). Rini suggests that even fake news with this aim is intended to deceive, because fake news can achieve virality only by finding credulous audiences willing to share it (2017: E45). Still, it is not clear from Rini's response that deceptive intent is part of the definition of fake news, rather than a feature fake news typically has. It seems possible for fake news to be non-accidentally misleading, in Fallis's sense, even if it is not intentionally deceptive. To the extent one finds the clickbait example compelling, one has reason to accept a relatively inclusive definition of fake news.

One might argue that fake news could in principle be generated and spread without human involvement—through bot networks, for instance—and hence does not require an intention to deceive. Arguably, disinformation might be generated in a similar fashion. However, I am inclined to think that whether the products of such a network would count as disinformation would depend on the intentions of the network's creators[6]. Denying the necessity of this intention would seem to allow for a conflation between misinformation and disinformation. While the former could be produced by a bot network established without manipulative intentions, the latter could not. As compared with disinformation, it is relatively plausible that fake news might be generated by a bot network without deceptive intentions. Such a suggestion would accord with the suggestion that fake news is sometimes akin to bullshit, rather than lies.

The arguments discussed in the preceding paragraphs offer some reasons to deny that fake news requires a deceptive intention[7]. While I will likewise deny that disinformation requires a deceptive intention, I do not do so for reasons paralleling those considered here. Ultimately, I maintain that disinformation requires an intention structurally like the intent to deceive. Thus, although the arguments developed below resemble in superficial respects the case for a relatively expansive definition of fake news, the way in which these arguments recommend expanding the relevant definitions differ significantly. This is not to say that fake news is never disinformation, but rather that fake news will only constitute disinformation when it is distributed with an intention along the lines I describe in Sect. 5.

In this section, I have sought to identify a widely shared commitment in accounts of disinformation. Commitment to a false belief account is not ubiquitous among commentators on disinformation. Indeed, we will see some existing alternatives below. However, given the prominence of false beliefs accounts, as demonstrated in this section, these accounts merit special attention.

[5] For a similar line of argument, see Jessica Pepp, Eliot Michaelson, and Rachel Katharine Sterken (2019: 73).

[6] Mona Simion (forthcoming) argues that a black-box AI that learns to systemically distribute false claims concerning the COVID-19 vaccines would thereby distribute disinformation, even in the absence of any deceptive intent on its part of the part of its human creators. Disagreement on this point may come down to intuition, but in my view it is unclear why this is a case of disinformation—as opposed to mere misinformation or fake news.

[7] For further reasons to this effect, see Grundmann (2020).

3 Disinformation, Disorientation, and Lies

Some commentators have found it instructive to investigate the concept of disinformation through comparison to lies. Fetzer (2004a): 231–232), for example, notes that the utility of comparing disinformation to lies is that both require an intention to mislead. Similarly, Fallis (2014: 138) suggests that lies are a kind of disinformation. However, insofar as lies are analogous to disinformation, the analogy cuts against false beliefs accounts of disinformation. Or so I now argue.

Standard philosophical analyses of the concept of lie take lying to require an intention to cause the audience to form a false belief[8]. Given this necessity, parallels between disinformation and lies would favor a false belief account of disinformation. However, it has been argued that such analyses fail to appreciate the possibility of lies that do not aim at causing false beliefs (Fallis, 2009b, 2010; Sorensen, 2007, 2010). Most relevant for present purposes are what Sorensen (2010) calls knowledge-lies. Sorensen illustrates the concept by reference to a scene in the film Spartacus, in which a series of Roman slaves claim to be Spartacus (2010: 608). Sorensen notes that, at least for those speakers later in the succession, it is not plausible that they intend to cause false beliefs. Instead, they intend to prevent the audience from learning which person is Spartacus. But they are lying nonetheless.

In general, knowledge-lies are intended to prevent knowledge without causing false belief (2010: 610). Supposing knowledge-lies are genuine lies, and that we have reason to accept a parallel between lies and disinformation, this parallel recommends against a false belief account of disinformation. Similarly, although Fallis commits to a false belief account of disinformation, he recognizes that lies can be intended to cause the abandonment or prevention of true beliefs (2014: 140–141). Assuming again that there is a parallel between lies and disinformation, such lies recommend a broader account of disinformation than the false belief account. Because neither I nor any other commentators in the literature have offered decisive reason to maintain a parallel between disinformation and lies, defenders of the false belief account may simply deny that the concepts share any relevant features[9]. However, such a strategy would surrender whatever motivation false belief accounts can receive from harmony with the more familiar concept of lies. Additionally, as I argue below, there is a class of disinformation that closely resembles knowledge-lies.

To recognize the existence of knowledge lies, one need not abandon the view that lies invariably involve deception. Lackey (2013) distinguishes between *deceit* and *deception*. In Lackey's telling, deceit involves the intent to cause false beliefs, while

[8] Lackey (2013) dates such analyses of lies back to at least Augustine, and cites Roderick Chisholm and Thomas Feehan (1977) and Williams (2002), among others, as more recent proponents.

[9] One might attempt to motivate such a denial by noting that one can share disinformation, but cannot lie, by credulously passing on false information that one believes to be true. However, the contrast between disinformation and lies is perhaps not so great as it initially appears. While lying plausibly involves—again at least in paradigmatic cases—and intention to deceive, it is often said that individuals spread lies even where such an intention is lacking. For instance, it has not been uncommon for commentators to accuse others of believing, and consequently spreading, Donald Trump's "Big Lie" concerning the integrity of the 2020 US presidential election. Such cases suggest a distinction between lying and spreading lies that may perhaps parallel a distinction between disinforming and spreading disinformation. Hence, the present point does not by itself warrant the denial of shared features between lies and disinformation.

deception involves only the intention to conceal information. Thus, on Lackey's view, knowledge lies involve deception even if they do not involve deceit. As we will now see, some forms of disinformation are deceptive in Lackey's sense, without involving deceit.

Some disinformation seems to aim not at producing false belief, but rather what Yochai Benkler, Robert Faris, and Hal Roberts call disorientation:

> Disorientation: a condition that some propaganda seeks to induce, in which the target population simply loses the ability to tell truth from falsehood or where to go for help in distinguishing between the two. (2018: 24).

In this passage, Benkler, Faris, and Roberts highlight a sought-after effect of some *propaganda*, rather than explicitly discussing disinformation. Elsewhere, however, the authors group together propaganda and disinformation as aiming at the manipulation of belief (2018: 6). To illustrate disinformation aimed at disorientation, consider the so-called "Firehose of Falsehood" model of propaganda[10] (Paul & Matthews, 2016). Observers of Russian propaganda have noticed that the apparent intention of many recent Russian disinformation operations is not to produce false beliefs, but is instead to pollute an audience's epistemic environment with such a confusing array of inconsistent information that the audience becomes unwilling to trust in anything at all (Giles, 2016). Consider the following example:

> When the Kremlin and its affiliated media outlets spat out outlandish stories about the downing of Malaysia Airlines Flight 17 over eastern Ukraine in July—reports that characterized the crash as everything from an assault by Ukrainian fighter jets following U.S. instructions, to an attempted NATO attack on Putin's private jet—they were trying not so much to convince viewers of any one version of events, but rather to leave them confused, paranoid, and passive—living in a Kremlin-controlled virtual reality that can no longer be mediated or debated by any appeal to 'truth.' (Pomerantsev 2014, emphasis added)

Given such cases, false belief accounts of disinformation appear too narrow[11]. The dissemination of multiple, mutually inconsistent narratives would be counterproductive to a disinformation campaign aimed at causing belief in some particular false narrative. However, the dissemination of such narratives can cause targets to lose trust in various sources of information[12] and, perhaps more perniciously, in their own

[10] Like Benkler, Faris, and Roberts in the passage above, Christopher Paul and Miriam Matthews sometimes use the term *propaganda*, rather than *disinformation*. However, the authors also extensively discuss the role of disinformation in producing confusion.

[11] Naomi Oreskes and Erik M. Conway (2010) provide a battery of further examples of what might be called disinformation. These cases typically involve industrial actors—the tobacco and fossil fuel industries, for some examples—promoting skepticism about the dangers of certain products. In these cases, the aim is not to encourage outright belief that the products in question are not harmful, but to generate doubt as to their dangers.

[12] The broader point here is that recognition of the existence of inaccurate information can lead to skepticism about sources of information more generally (cf. Fallis, 2004: 465).

abilities to tell truth from falsehood. In short, some forms of disinformation aim to manipulate not by instilling falsehood, but by preventing their targets from believing the truth (Rini, 2021). Notably, the phenomenon just described, which I have suggested is a form of disinformation, closely parallels knowledge-lies. Just as Spartacus's companions aim most plausibly at preventing true beliefs, rather than causing false beliefs, the dissemination of mutually inconsistent narratives is best understood as aimed at obscuring the truth, rather than causing false beliefs.

That disinformation sometimes aims at preventing true beliefs accords well with Lackey's suggestion that the aim of deceptive acts is sometimes to conceal information. Moreover, some existing discussions of disinformation hint at a similarly inclusive notion. While Fetzer sometimes seems to favor a false belief account of disinformation (2004b: 228), he suggests elsewhere that disinformation sometimes aims to produce confusion (2004b: 231). Similarly, while Fallis explicitly supports a false belief approach in those passages cited above, he does so alongside the recognition that deception sometimes aims at hiding the truth, rather than showing the false (2014: 140–141). Elsewhere, Fallis floats, without endorsing, the proposal that disinformation functions to promote false beliefs or prevent true beliefs (2015: 420). There are, in short, antecedents to the relatively broad approach to disinformation advocated here in the existing literature.

I have thus far argued that disinformation need not aim at producing false beliefs and indeed need not have the production of false beliefs as its function. One might take the upshot of the preceding argument to be that the function of disinformation is to produce false beliefs or to prevent true beliefs. Such a conclusion would be significant not only for the sake of correcting misapprehensions about the nature of disinformation, but also in light of the practical consequences of the more encompassing understanding of disinformation advocated here. Before considering these practical considerations, I make the case for a further broadening of our understanding of disinformation.

4 Disinformation and Sub-doxastic States

In this section and the next, I argue that the aims of disinformation go beyond influence on beliefs—either in the form of causing false beliefs or preventing true beliefs. To make this argument, I begin by considering some important challenges to the attempt to explain human behavior in terms of belief-desire psychology. Ultimately, I will argue that, because important human behaviors are shaped by what I will call *sub-doxastic states*, an account of disinformation should recognize that disinformation can target such states.

As an illustration of sub-doxastic states, consider the following example first discussed by Gendler (2008a). The Grand Canyon Skywalk is a seventy-foot-long glass walkway protruding into the canyon. Visitors on the walkway, which is carefully protected from scratches and other damage that might compromise its transparency, can look beneath them to see the canyon floor roughly 4,000 feet down. As Gendler reports, some visitors to the walkway cannot bring themselves to walk over it. This is true even as these visitors have excellent evidence, including the ability to witness others walking along the walkway, of the structure's integrity. Indeed, one might sincerely report a belief in the structure's integrity even as one cannot bring oneself to walk upon it. According to Gendler, such mismatches between belief and behavior indicate a challenge for the program of explaining behavior in terms of belief-desire psychology.

The upshot, according to Gendler, is the need to introduce a distinct kind of mental state, the *alief*, to account for certain behaviors. As Gendler describes them, aliefs are "*a*ssociative, *a*utomatic, and *a*rational…And they are typically also *a*ffect-laden and *a*ction generating" (2008a: 641). It should be noted from the outset that the need to introduce alief to account for aspects of human behavior is controversial (Mandelbaum, 2013). Yet it is not controversial that cases like that involving the Grand Canyon Walkway are real and illustrate a class of human behavior with significant real-world impacts. To illustrate the latter point, let us consider another variety of behavior that has received substantial attention in recent decades.

There now exists a large body of psychological work suggesting that one factor in the perpetuation of inequalities between social groups is the existence of so-called implicit biases (Brownstein & Saul, 2016). Implicit biases are often, though not universally (Gawronski & Bodenhausen, 2006; Hahn & Gawronski, 2014), thought to be unconscious and beyond the agent's control. Implicit biases, like Gendler's aliefs, are usually taken to be associative states. Implicit biases may, for instance, involve associations between certain racial groups and danger (Correll et al., 2002). Indeed, Gendler proposes to understand implicit bias in terms of alief (2008b).

Implicit biases are pernicious, in part, because such biases may persist even in those subjects that resist them. For example, implicit racial biases are observable even in those subjects who explicitly disavow racism (Dovidio & Gaertner, 2004). This combination of attitudes is sometimes called *aversive racism*. Recognition of aversive racism is concerning, in part, because it suggests that even those who sincerely commit themselves to anti-racism may nonetheless perpetuate racial inequalities because of their implicit biases. Such biases have, for example, been implicated in inequalities in hiring practices (Bertrand & Mullainathan, 2004) and police use of force (Correll et al., 2002).

To say that implicit biases resist correction is not to say that such biases are unresponsive to experience. Empirical evidence suggests that even brief interventions may amplify or dampen expression of implicit biases (Foroni & Mayr, 2005; Wittenbrink et al., 2001). Moreover, a large body of empirical work suggests that such attitudes effectively mirror the environment (Dasgupta, 2013). Thus, while implicit biases are likely to develop in environments in which the objects of bias are disvalued, changes in the environment can reduce or eliminate biases (Huebner, 2016).

The picture that emerges from consideration of aliefs and implicit biases accords with the two-tiered picture of the mind suggested by dual-process theory. In addition to the level of beliefs, there appears to be a further level of mental activity to which relatively automatic, non-rational, associative states are central. The existence of this second level is most evident when the states involved are mismatched with the subject's doxastic states, as in cases of aversive racism and sexism and in cases like the Grand Canyon Walkway. While these sub-doxastic associative states are usually thought to be resistant to evidence, I have noted above that they can be influenced by exposure to certain kinds of content. Recognizing this two-tiered picture is a first step toward making evident a further shortcoming of the accounts of disinformation considered in Sect. 2. Moreover, this picture suggests that these accounts cannot be salvaged merely by broadening focus from the generation of false beliefs to also include the role of disinformation in preventing true beliefs. Such an account would fail to appreciate the manipulability of the mind by interventions that do not target beliefs. I now argue that an adequate account of disinformation should allow that the function of disinformation may be to manipulate sub-doxastic associative states.

5 What is Disinformation?

I have thus far argued that accounts of disinformation centered on the generation of false beliefs are too narrow, and I have drawn on empirical work to suggest that human behavior is party shaped by sub-doxastic associative states. I have, however, not yet shown that attention to such states has a role to play in the correct understanding of disinformation. To begin to make the case, let us consider a second family of accounts of disinformation, beyond that considered in Sect. 2.

In a major report on disinformation authored for the European Commission, disinformation is defined as:

> [F]alse, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit. (de Cock Buning 2018: 35)

Howard Tumber and Silvio Waisbord take disinformation to refer to:

> [D]eliberate attempts to sow confusion and misinformation among the public, with the purpose of political gain by a range of public, private, and social actors. (2021: 1)

W. Lance Bennett and Steven Livingston provide the following provisional definition of disinformation:

> [I]ntentional falsehoods spread as news stories or simulated documentary formats to advance political goals. (2018: 124)

These definitions are not equivalent. Moreover, Bennett and Livingston ultimately adopt an approach to understanding disinformation more in line with the false belief accounts considered in Sect. 2. However, the definitions given here draw attention to an important feature of disinformation that is neglected by the false belief accounts—namely that the ultimate function of disinformation is not to affect individuals' attitudes. Rather, disinformation has the function, at least typically in virtue of the aim of its practitioners, of influencing the target's behavior. As Bennett and Livingston suggest, the desired changes in behavior are often political in nature, but may also or instead be centered on financial profit.

Attending to the ultimate function of disinformation helps to clarify why an adequate definition of disinformation should include reference to sub-doxastic states. Disinformation aims to influence the behavior of its targets, but this aim need not be carried out through manipulation of beliefs. It may instead be carried out through manipulation of sub-doxastic states. Indeed, the relative insensitivity of sub-doxastic states to counterevidence that even the subject recognizes as such may make such states ideal targets from the perspective of the disinformation agent. For example, a malign actor seeking to perpetuate racist structures might in principle further this aim through the dissemination of material intended to influence individual's sub-doxastic states. As the case of aversive racism illustrates, this goal might be accomplished even if the targets' doxastic states are unaffected. Extant empirical data on the use of disinformation to manipulate via changes to the target's sub-doxastic states is scant, but some recent research suggests that disinformation can influence behavior in this way (Bastick, 2021).

Consider some additional examples. In the context of a political contest, it may be beneficial to make potential voters believe that one's opponent is guilty of impropriety. But one may manipulate voter behavior without influencing beliefs. For example, one might accuse one's opponent of offenses so outrageous that few not already disposed to dislike that candidate would believe them. One might not thereby change many explicit beliefs, but one might nonetheless cause one's audience to associate the opponent with those allegations. This technique is closely related to what is called in Russian the "rotten herring" technique, which is usually understood to succeed when allegations dog a target in public (Perianova, 2019: 160), but the present contention is that such allegations will succeed when they are internalized, even if not at the level of belief[13]. For a final example, consider the form that a great deal of misinformation concerning COVID-19 vaccines has taken. Much of this misinformation is in the form of videos that purport to show recipients of vaccines encountering severe side effects, including uncontrollable shaking and difficulty walking (Vulpicelli, 2021). There is no reason to expect that these graphic videos would be more convincing than textual claims as to the supposed danger posed by the vaccines. But videos of this sort are psychologically gripping, and can be expected to cause vaccine hesitancy even among those that do not recognize a rational basis for hesitancy. Indeed, earlier studies have shown that vaccine hesitancy outside of the context of the COVID-19 pandemic are closely tied to feelings (Tomljenovic, Bubic & Erceg, 2020) and that anti-vaccination websites are effective at producing such negative feelings (Betsch et al., 2010).

With the preceding considerations and examples in mind, I suggest the following account of disinformation:

Disinformation Content intended to manipulate the behavior of an audience by causing the audience to form counter-normative beliefs or belief-like states or by preventing the audience from holding normative beliefs or belief-like states.

[13] In a similar vein, Stanley (2015: ch. 4) writes that some linguistic propaganda works by associating words with problematic stereotypes or images. Thanks to an anonymous referee for drawing my attention to this parallel.

This account requires some unpacking. First, the audience whose beliefs or belief-like states are affected may be one individual, multiple separate individuals, or a group[14]. Some philosophers have argued that groups may hold beliefs not held by any of their members (Bird, 2018; Gilbert, 1987). In light of this possibility, it is consistent with the present account that disinformation sometimes targets group beliefs, without targeting the beliefs of any individual. Second, in line with Fallis's (2014) point that deception may be intended to cause the abandonment of or failure to form true beliefs, the present account allows that disinformation might aim to lead audiences to abandon or simply fail to form true beliefs or normative belief-like states. Third, to say that disinformation is intended to manipulate is not to say that all purveyors of disinformation intend to manipulate. In some cases, disinformation may originate at the hands of manipulative parties, but be spread in large part by true believers. Fourth, I use the general term 'content' here in recognition of the fact that disinformation can take the form of verbal assertions, statistics, pictures, audio or video recordings, written statements, internet memes, and so on (Fallis, 2014). Finally, it is worth emphasizing that this definition identifies disinformation as a kind of content, without specifying that such content must be false, misleading, or anything of the sort. This is because even entirely accurate content might constitute disinformation. For example, a nativist political actor might repeatedly raise statistics concerning the number of crimes committed by immigrants. Even if these statistics are accurate, and indeed even if they neither mislead nor are intended to mislead the audience, the repetition of these statistics might lead the audience to form or strengthen counter-normative psychological associations between immigrants and crime[15].

The preceding point makes salient the need to clarify what it is for an associative state to be normative or counter-normative. To grasp the issue, note that there is a well-established literature on the norms for belief. However, the norms widely thought to apply to belief—truth (Wedgwood, 2013; Whiting, 2010) or perhaps knowledge (Williamson, 2000)—are at least potentially applicable to belief in virtue of its propositional object. But associative states are non-propositional, and hence cannot be assessed in terms of truth or knowledge norms (cf. Mandelbaum, 2013: 202).

It is possible to describe certain senses in which associative states might be normative or counter-normative. For example, we might plausibly regard associative states as counter-normative insofar as they tend to produce counter-normative behavior. For example, an associative state linking immigrants to crime might be counter-normative insofar as it promotes prejudicial behavior toward immigrants. Likewise, it is evident in much of the existing literature on implicit bias that such biases are counter-normative at least in the sense that they tend to produce counter-normative behavior. As noted above, implicit biases may for example result in unfair hiring decisions. Yet assessing the normative status of associative states strictly in terms of behavioral consequences is problematic. Some beliefs, like some associative states,

[14] Thanks to an anonymous referee for recommending that I emphasize this point.

[15] The example is not merely speculative. Even brief exposures to rightwing populist political posters linking immigrants to crime have been found to strengthen implicit associations between these concepts (Arendt, Marquart & Matthes, 2015; Matthes and Schmuck, 2017).

may tend to result in problematic behavior. However, as we have seen, this is not the only sense in which beliefs themselves may be counter-normative. Beliefs have distinctive norms, independent of their behavioral consequences. It is these norms that figure in the definition of disinformation given above. Specifically, in the case of belief, disinformation works through promotion of false belief and prevention of true belief. The real task, then, is to state normative principles for associative states that parallel the behavior-independent norms of belief.

Some might contend that there can be no such norms—that the existence of an ethics within a domain of human activity is contingent on human control over that activity—something lacking in the case of associative states. This response is excessively pessimistic. It has been argued that the ethics of belief is not contingent on the voluntariness of belief (Feldman, 1988; Steup, 2000). Moreover, it is not clear that ordinary humans completely lack control over their sub-doxastic associative states (Frankish, 2016). This control may be indirect, but so too, it has been argued, is control over beliefs (Heil, 1983; Leon, 2002; Price, 1954). What is more, we may locate a thin sense of normativity operative even where control is absent. Intuitively, the reading of a broken thermometer is, or is likely to be, counter-normative. Thus, we should not abandon too quickly the possibility of an ethics of associative states.

Still, such an ethics is elusive. It may be thought that norms for associative states can be constructed to closely parallel those for beliefs by considering the default interface between associative and propositional thought. Bertram Gawronski and Galen V. Bodenhausen write that "the default mode of propositional thinking is an affirmation of momentarily activated associations" (2006: 694). One might thus suppose that an associative state is normative just in case its default propositional counterpart is normative. For example, the association of fire with warmth is normative just in case the belief that *fire is warm* is true. Yet this suggestion faces immediate difficulty. Consider again the association of immigrants with criminals. The most apparent propositional counterpart to this association is something along the lines of *immigrants are criminals*. Such a proposition has an air of falsity, and thus counter-normativity to it, but only insofar as one resolves the ambiguity in the claim to assert that *all immigrants are criminals*, *immigrants are more likely than non-immigrants to be criminals*, or something of the like. But it is not clear why some proposition along these lines, rather than some true proposition along the lines of *some immigrants are criminals*, is the proposition from which the normative status of the corresponding associative state is to be derived[16]. In short, the attempt to determine the normative

[16] As an anonymous referee has suggested, the challenge here arises from the fact that the apparent propositional counterparts of associative states will typically be expressible as generics. These are expressions that make generalizations without any explicit quantifiers. Various proposals concerning the truth conditions of generics have been advanced (Cohen, 2013; Leslie, 2007; Liebesman, 2011; Pelletier & Asher, 1997), but thus far none has achieved consensus. This is despite the fact that we competently use and understand generics, even from a young age (Leslie, 2007). It is thus unsurprising that we may regard the apparent propositional counterparts of certain associative states as problematic, even without being able to state clear truth conditions for those propositions. As a final note on generics, it is worth mentioning in this context that Sarah-Jane Leslie has taken the mode of generalization involved in processing generics to be implicated in the formation of prejudices (2017). If so, there is reason to think that some disinformation might achieve its aim by exploiting this mode of generalization.

status of a given associative state from that of a corresponding proposition flounders on the lack of a rule for translating between associations and propositions.

The failure of this proposal should not make us despair of the availability of an ethics for associative states. Perhaps normativity for associative states can be defined in relation to the associative states that would be had by some ideal reasoner in an ideal informational environment. Perhaps the ambiguity issue raised in connection with the first proposal can be resolved by attending to the co-activation of concepts alongside a broader network of associated concepts, sufficient to pick out at least a narrower range of counterpart propositions. Or perhaps the normativity of any given association can only be assessed in comparison with the subject's other associations. For example, one who strongly associates immigrants, but not non-immigrants, with crime appears thereby to err. I will not attempt to settle the matter here, preferring instead to rely on the intuitive judgment that the associative states apparently intended by some disinformation—for example the association of immigrants with crime—are counter-normative. Moreover, some associations plausibly suppressed by disinformation—associations of immigrants with humanity, family, and virtue, for example—are intuitively normative.

To conclude this section, let us consider a pair of related objections. Unlike the false belief accounts of disinformation, the account offered here recognizes attempts to manipulate audiences by way of preventing true beliefs or influencing sub-doxastic associative states as involving disinformation. But why should we think such attempts involve disinformation? This objection might be compounded by the concern that, unlike everyday concepts like *knowledge*, we lack the requisite pre-theoretic intuitions to discern the boundaries of disinformation. One might thus maintain that the concept of disinformation is more suitable for engineering than analysis.

There are at least three responses that one might give to these objections. While I favor the third, I do not deny that the alternatives might be fruitfully pursued. First, one might largely concede to the objections but argue that either disinformation is inclusive in the way specified here *or* focusing on disinformation is too narrow and misses other important causes of epistemic dysfunction. Second, one might recognize the importance of the phenomena described and thus argue for engineering *disinformation* in an inclusive fashion. Finally, one might argue that there is good reason to analyze disinformation in the inclusive way described here. In addition to the aforementioned point that the present account unites the belief-focused accounts of disinformation with the behavioral-manipulation accounts, the present account reflects a realistic concern with the priorities of purveyors of disinformation. As the accounts discussed at the beginning of this section suggest, disinformation seems to aim at manipulating behavior. Yet, while it is plausible that would-be manipulators distinguish between disinforming and coercing, it is dubious that such manipulators generally recognize a distinction between belief states and sub-doxastic associative states. Such a distinction is familiar to philosophers and cognitive scientists, but is not part of folk psychology. Thus, it is more plausible that would-be manipulators sometimes aim to manipulate their targets by counter-normatively influencing a generic class of mental states that includes beliefs and belief-like states. The present definition thus reflects a realistic picture of the likely aims of agents of disinformation.

6 Beyond Belief: The Threat of Disinformation

I have argued that doxastic accounts of disinformation are too narrow insofar as they fail to appreciate that disinformation may be intended to manipulate without interfering with the target's doxastic states. Recognizing this point is, I now argue, key to grasping the full extent of the threat that disinformation poses. If one remains skeptical of the inclusive account of disinformation offered here, the remarks to follow might instead be taken to motivate greater attention to phenomena other than disinformation or to motivate a particular way of engineering the *disinformation* concept.

If disinformation is construed as aiming to produce false beliefs, the task of combatting disinformation appears relatively straightforward. On the face of things, the effectiveness of disinformation, so construed, could be substantially mitigated by exposure to competing or contextualizing truths. Some commentators have suggested that the chief challenge to combating disinformation, so understood, is to win attention to the truth and its promoters (Smith & Wanless 2020). This suggestion is perhaps overly optimistic in light of certain features of the social epistemic environment. Consider C. Thi Nguyen's (2020) distinction between epistemic bubbles and echo chambers. Both phenomena plausibly play a role in the formation and preservation of false beliefs. Epistemic bubbles, which Nguyen suggests are characterized by a lack of connectivity to certain sources, might be addressed by drawing attention to the truth and its promoters. But echo chambers, which are characterized in part by distrust of outsiders, cannot be so easily dismantled by exposure to the truth and trustworthy sources[17]. Indeed, empirical evidence for the so-called backfire effect appears to illustrate that exposure to competing perspectives and apparent counterevidence may deepen one's ideological commitments (Nyhan & Reifler, 2010; Nyhan et al., 2013). However, the role of the backfire effect in preserving false beliefs is often overstated (Guess & Coppock 2020; Nyhan, 2021), and substantial evidence indicates that false beliefs, even concerning politically significant matters, can be at least temporarily corrected through exposure to accurate information (Nyhan et al., 2020; Wood & Porter, 2019). None of this is to deny that echo chambers are real and may insulate false beliefs against correction in some cases, but these findings suggest that exposure to truth is an effective antidote to false belief in at least some cases.

The prospects for combating disinformation with truth dim further as we broaden our focus beyond misleading disinformation. As we have seen, disinformation may aim at the prevention of true belief. Attempts to meet disinformation with truth may contribute to the success of disinformation, so construed. After all, any such effort makes salient the epistemic vulnerability to which reliance on others exposes us. In short, correcting for misleading information makes the existence of such information salient, thereby discouraging belief. In this way, disinformation may co-opt attempts to combat it into furthering its own aims.

Disinformation that aims at manipulating targets via their sub-doxastic states is likewise unlikely to be effectively countered through the dissemination of truth alone. Gendler, as we have seen, describes aliefs as arational. More generally, counter-nor-

[17] For an in-depth study of a political echo chamber in the American context, see Kathleen Hall Jamieson and Joseph N. Cappella (2008).

mative associative states may persist despite the subject's sincere attempts to correct them. In short, even if truth can mitigate false belief, there is less reason to suppose that it will mitigate counter-normative sub-doxastic states. Just as a pedestrian on the Grand Canyon Walkway might quake with fear despite decisive evidence as to the structural integrity of the walkway, the target of disinformation might exhibit counter-normative sub-doxastic states, even while explicitly rejecting the legitimacy of the disinformational content.

While the preceding remarks suggest a highly pessimistic outlook on the threat of disinformation and the prospects for effectively countering that threat, let us conclude this section on a relatively optimistic note. The account of disinformation given here suggests three ways in which disinformation may affect human agency. Misleading disinformation aims to pervert human agency, shaping human action by causing false beliefs. As we have seen, alternative forms of disinformation aim to prevent targets from forming true beliefs, thereby reducing their agency. Finally, some forms of disinformation aim to manipulate targets' behavior via their sub-doxastic states, thereby sidestepping targets' agency. This final form of disinformation may seem the most pernicious, insofar as counter-normative sub-doxastic attitudes can persist even in the face of explicit disavowal on the part of their subjects. Yet there are two reasons for optimism on this score. First, in cases of tension between a subject's normative doxastic and counter-normative sub-doxastic states, the subject may endeavor to ensure that her behavior is guided by the former, and not the latter. Just as an aversive racist might opt to anonymize resumes to prevent biased associations with names from influencing decisions, targets of disinformation might take steps to avoid manipulation via sub-doxastic states. For example, a voter concerned about political disinformation might opt to vote strictly based on candidates' stances, rather than on feelings toward candidates. Strategies of this type aim at mitigating the behavioral effects of counter-normative associations, rather than at reducing counter-normative associations themselves. An individual can thereby defend herself from disinformation that would otherwise sidestep her agency. However, such strategies may be impractical, especially in those contexts—such as police use of force decisions—which call for immediate action.

A second cause for relative optimism highlights the possibility of reducing counter-normative associations themselves. Empirical evidence suggests that sub-doxastic attitudes are malleable, effectively mirroring features of the subject's environment. Insofar as this environment is polluted by disinformation and related phenomena, subjects are likely to form counter-normative sub-doxastic attitudes, and to maintain these even despite conscious rejection. However, the malleability of sub-doxastic attitudes suggests that improvements to such states are possible through improvement of the epistemic environment. Given that sub-doxastic states are arational, the needed improvements do not merely consist in making available information to debunk, disprove, or contextualize disinformation. Instead, we may expect that the deleterious effects of disinformation on sub-doxastic states can only be successfully mitigated through large-scale and repetitive protective and corrective interventions in the epistemic environment. For example, the mere availability of information to debunk anti-vaccine disinformation is unlikely to mitigate certain individuals' negative associations with vaccines (Tomljenovic et al., 2020: 549). Just as such associations are likely to form in the presence of repeated exposure to various forms of disinformation linking vaccines to negative outcomes, governmental overreach, and the like (Betsch et al., 2010), it is to be expected that such associations are most effectively mitigated by repeated exposure to information linking vaccines to positive outcomes. In short, effectively countering the effects of disinformation on sub-doxastic states is likely to require interventions on the broader epistemic environment that support normative associations. The democratization of control over this environment due to the emergence of social media and related developments suggests that the improvement of this environment is a project in which we may all partake.

7 Concluding Remarks

I have argued that the understanding of disinformation as aiming or functioning to produce false beliefs is untenably narrow. Some styles of disinformation may instead aim at preventing true beliefs. But to understand disinformation strictly in terms of effects on belief would be to overintellectualize it. As cases like the hesitant pedestrian on the glass walkway demonstrate, human behavior cannot be explained purely in terms of beliefs and desires. For this reason, human behavior can be manipulated by mechanisms that bypass belief and desire. Disinformation may operate, and indeed may be most effective, at the level of gut feelings. This broader understanding of disinformation brings into focus the inadequacy of certain interventions against disinformation and the importance of sophisticated responses to the threat that recognize the complexity of human behavior and its corresponding vulnerability to manipulation.

Acknowledgements Thanks to audiences at the 2022 Science and the Art of Simulation and German Analytic Philosophy conferences, as well as several anonymous referees, for their feedback on earlier versions of this paper. This paper was 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. The sole responsibility for the content of this publication lies with the author.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Arendt, F., Marquart, F., & Matthes, J. (2015). Effects of right-wing populist political advertising on implicit and explicit stereotypes. *Journal of Media Psychology*, *27*(4), 178–189.

Bastick (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior*, *116*, 106633.

Behavioral and Brain Sciences, 37(1): 28–9.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in american politics*. Oxford University Press.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*, 991–1013.

Betsch, C., Renkewitz, F., Betsch, T., & Ulshöfer, C. (2010). The influence of vaccine-critical websites on perceiving vaccination risks. *Journal of Health Psychology*, *15*(3), 446–455.

Bird, A. (2018). Group belief and knowledge. In M. Fricker, P. J. Graham, D. Henderson, & N. J. J. L. Pedersen (Eds.), *The Routledge Handbook of Social Epistemology* (pp. 274–283). NY: Routledge.

Brownstein, M., & Saul, J. (2016). Introduction. In M. Brownstein, & J. Saul (Eds.), *Implicit Bias and Philosophy – volume 1: Metaphysics and Epistemology* (pp. 1–19). NY: Oxford University Press.

Cassam, Q. (2019). *Conspiracy theories*. Polity Press.

Chisholm, R. M., & Feehan, T. D. (1977). The intent to deceive. *Journal of Philosophy*, *74*(3), 143–159.

Chivvis (2016). Understanding russian hybrid warfare. *Rand Corporation*. 1–10.

Cohen, A. (2013). No quantification without reinterpretation. In A. Mari, C. Beyssade, & F. Del Prete (Eds.), *Genericity* (pp. 334–351). Oxford: Oxford University Press.

Cooper, G. (2021). Populist rhetoric and media misinformation in the 2016 UK Brexit referendum. In H. Tumber, & S. Waisbord (Eds.), *The Routledge Companion to Media Disinformation and Populism* (pp. 397–410). NY: Routledge.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology, Vol. 47* (pp. 233–279). Elsevier Academic Press.

De Cock Buning, M. (2018). *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union.

Dovidio & Gaertner, Dovidio, J. F., & Gaertner, S. L. (2004). (2004). Aversive racism. In M.P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 36* (pp. 1–52). Elsevier Academic Press.

Fallis, D. (2004). On verifying the accuracy of information: Philosophical perspectives. *Library Trends*, *52*, 463–487.

Fallis, D. (2009a). A conceptual analysis of disinformation. Paper presented at the fourth annual iConference at University of North Carolina, Chapel Hill. https://www.ideals.illinois.edu/handle/2142/15201/browse.

Fallis, D. (2009b). What is lying? *The Journal of Philosophy*, *106*(1), 29–56.

Fallis, D. (2010). Lying and deception. *Philosopher's Imprint*, *10*(11), 1–22.

Fallis, D. (2014). The varieties of disinformation. In L. Floridi, & P. Illari (Eds.), *The philosophy of Information Quality* (pp. 135–161). NY: Springer.

Fallis, D. (2015). What is disinformation? *Library Trends*, *63*(3), 401–426.

Feldman, R. (1988). Epistemic obligations. *Philosophical Perspectives*, *2*, 235–256.

Fetzer, J. (2004a). Disinformation: The use of false information. *Minds and Machines*, *14*, 231–240.

Fetzer, J. (2004b). Information: Does it have to be true? *Minds and Machines*, *14*, 223–229.

Floridi, L. (2011). *The philosophy of information*. Oxford: Oxford University Press.

Floridi, L. (2012). Steps forward in the philosophy of information. *Ethics & Politics*, *14*(1), 304–310.

Foroni, F., & Mayr, U. (2005). The power of a story: New, automatic associations from a single reading of a short scenario. *Psychonomic Bulletin & Review*, *12*(1), 139–144.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton: Princeton University Press.

Frankish, K. (2016). Playing double: Implicit bias, dual levels, and self-control. In M. Brownstein, & J. Saul (Eds.), *Implicit Bias and Philosophy – volume 1: Metaphysics and Epistemology* (pp. 23–46). NY: Oxford University Press.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731.

Gendler, T. S. (2008a). Alief and belief. *Journal of Philosophy*, *105*(10), 634–663.

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, *23*(5), 552–585.

Gilbert, M. (1987). Modelling collective belief. *Synthese*, *73*(1), 185–204.

Giles, K. (2016). *Handbook of Russian Information Warfare*. Rome: NATO Defense College.

Glenski, V., & Kumar (2020). User engagement with digital deception. In K. Shu, S. Wang, D. Lee, & H. Liu (Eds.), *Disinformation, misinformation, and fake news in Social Media Emerging Research Challenges and Opportunities* (pp. 39–61). Springer.

Grundmann, T. (2020). Fake news: The case for a purely consumer-oriented explication. *Inquiry: A Journal Of Medical Care Organization, Provision And Financing*. https://doi.org/10.1080/0020174X.2020.1813195.

Guess, A., & Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, *50*(4), 1497–1515.

Hahn, A., & Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes?.

Harris, K. (2022). Conspiracy theories, populism, and epistemic autonomy. *Journal of the American Philosophical Association*, 1–16. https://doi.org/10.1017/apa.2021.44.

Heil, J. (1983). Doxastic agency. *Philosophical Studies*, *43*(3), 355–364.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein, & J. Saul (Eds.), *Implicit Bias and Philosophy – volume 1: Metaphysics and Epistemology* (pp. 47–79). NY: Oxford University Press.

Jaster, R., & Lanius, D. (2021). Speaking of fake news: Definitions and dimensions. In S. Bernecker, A. K. Flowerree, & T. Grundmann (Eds.), *The epistemology of fake news* (pp. 17–45). Oxford: Oxford University Press.

Lackey, J. (2013). Lies and deception: An unhappy divorce. *Analysis*, *72*(2), 236–248.

Leon, M. (2002). Responsible believers. *Monist*, *85*(3), 421–435.

Leslie, S. J. (2007). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1–48.

Leslie, S. J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *Journal of Philosophy*, *114*(8), 393–421.

Liebesman, D. (2011). Simple generics. *Noûs*, *45*(3), 409–442.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behavior*, *5*(3), 337–348.

Mahon, J. E. (2007). A definition of deceiving. *International Journal of Applied Philosophy*, *21*(2), 181–194.

Mahon, J. E. (2015). The definition of lying and deception. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/lying-definition/.

Mandelbaum, E. (2013). Against alief. *Philosophical Studies*, *165*(1), 197–211.

Matthes, J., & Schmuck, D. (2017). The effects of anti-immigrant right-wing populist ads on implicit and explicit attitudes: A moderated mediation model. *Communication Research*, *44*(4), 556–581.

McIntyre, L. (2018). *Post-Truth*. MIT Press.

Mueller, R. (2019). *Report on the investigation into russian interference in the 2016 presidential election* (1 vol.). U.S. Department of justice.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161.

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15).

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330.

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, *51*(2), 127–132.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*, 939–960.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from Tobacco smoke to global warming*. New York: Bloomsbury Press.

Paul, C., & Matthews, M. (2016). The russian “firehose of falsehood” propaganda model: Why it might work and options to counter it. *RAND Corporation*. 1–16.

Pelletier, F., & Asher, N. (1997). Generics and defaults. In. In van J. Benthem, A. ter, & Meulen (Eds.), *Handbook of Logic and Language* (pp. 1125–1179). Cambridge: MIT Press.

Pepp, J., Michaelson, E., & Sterken, R. K. (2019). What’s new about fake news? *Journal of Ethics and Social Philosophy*, *16*(2), 67–94.

Perianova, I. (2019). *A Mashup World: Hybrids, Crossovers and Post-Reality*. Cambridge Scholars Publishing.

Pomerantsev, P. (2014). Russia and the menace of unreality. *The Atlantic*. 9 September 2014. https://www.theatlantic.com/international/archive/2014/09/russia-putin-revolutionizing-information-warfare/379880/.

Price, H. H. (1954). Belief and will. *Proceedings of the Aristotelian Society*, *28*(1), 1–26.

Rini, R. (2017). Fake news and partisanship. *Kennedy Institute of Ethics Journal*, *27*(2-Supplement), E43–E64.

Rini, R. (2019). Social media disinformation and the security threat to democratic legitimacy. NATO Association of Canada: *Disinformation and Digital Democracies in the 21st Century*.10–14.

Rini, R. (2021). Weaponized skepticism: An analysis of social media deception as applied political epistemology. In E. Edenberg, & M. Hannon (Eds.), *Political epistemology* (pp. 31–48). Oxford: Oxford University Press.

Saul, J. (2012). *Lying, Misleading, and what is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford University Press.

Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In K. Shu, S. Wang, D. Lee, & H. Liu (Eds.), *Disinformation, misinformation, and fake news in Social Media Emerging Research Challenges and Opportunities* (pp. 1–19). Springer.

Simion, M. (forthcoming) (Ed.). Knowledge and disinformation. *Episteme*.

Skyrms, B. (2010). *Signals: Evolution, Learning, and information*. Oxford: Oxford University Press.

Smith & Wanless (2020). Unmasking the truth: Public health experts, the coronavirus, and the raucous marketplace of ideas. Carnegie Endowment for International Peace: *Partnership for Countering Influence Operations*. 1–27.

Søe, S. O. (2021). A unified account of information, misinformation, and disinformation. *Synthese*, *198*(6), 5929–5949.

Sorensen, R. (2007). Bald-faced lies! Lying without the intent to deceive. *Pacific Philosophical Quarterly*, *88*(2), 251–264.

Sorensen, R. (2010). Knowledge-lies. *Analysis*, *70*(4), 608–615.

Stanley, J. (2015). *How Propaganda Works*. Princeton University Press.

Steup, M. (2000). Doxastic Voluntarism and Epistemic Deontology. *Acta Analytica*, *15*(1), 25–56.

Tagliabue, F., Galassi, L., & Mariani, P. (2020). The “pandemic” of disinformation in COVID-19. SN comprehensive clinical medicine, 1–3. https://doi.org/10.1007/s42399-020-00439-1.

Tomljenovic, H., Bubic, A., & Erceg, N. (2020). It just doesn’t feel right – the relevance of emotions and intuition for parental vaccine conspiracy beliefs and vaccination uptake. *Psychology & Health*, *35*(5), 538–554.

Tumber, H., & Waisbord, S. (2021). Introduction. In H. Tumber, & S. Waisbord (Eds.), *The Routledge Companion to Media Disinformation and Populism* (pp. 1–12). New York: Routledge.

Vulpicelli, G. M. (2021). They claimed the Covid-19 vaccine made them ill. Then they went viral. *Wired*. 23 January 2021. https://www.wired.co.uk/article/covid-vaccine-misinformation-facebook.

Wedgwood, R. (2013). The aim of belief. *Philosophical Perspectives*, *16*, 267–297.

Whiting, D. (2010). Nothing but the truth: On the norms and aim of belief. In T. Chan (Ed.), *The Aim of Belief* (pp. 184–203). New York: Oxford University Press.

Williams, B. (2002). *Truth and truthfulness: An essay in Genealogy*. Princeton: Princeton University Press.

Williamson, T. (2000). *Knowledge and its limits*. New York: Oxford University Press.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes’ steadfast factual adherence. *Political Behavior*, *41*, 135–163.