Misinformation, Content Moderation, and Epistemology

This book argues that misinformation poses a multifaceted threat to knowledge, while arguing that some forms of content moderation risk exacerbating these threats. It proposes alternative forms of content moderation that aim to address this complexity while enhancing human epistemic agency.

The proliferation of fake news, false conspiracy theories, and other forms of misinformation on the internet and especially social media is widely recognized as a threat to individual knowledge and, consequently, to collective deliberation and democracy itself. This book argues that misinformation presents a three-pronged threat to knowledge. While researchers often focus on the role of misinformation in causing false beliefs, this deceptive potential of misinformation exists alongside the potential to suppress trust and to distort the perception of evidence. Recognizing the multifaceted nature of this threat is essential to the development of effective measures to mitigate the harms associated with misinformation. The book weaves together work in analytic epistemology with emerging empirical work in other disciplines to offer novel insights into the threats posed by misinformation. Additionally, it breaks new ground by systematically assessing different forms of content moderation from the perspective of epistemology.

Misinformation, Content Moderation, and Epistemology will appeal to philosophers working in applied and social epistemology, as well as scholars and advanced students in disciplines such as communication studies, political science, and social psychology who are researching misinformation.

Keith Raymond Harris is a postdoctoral researcher in philosophy at Ruhr-Universität Bochum, where he specializes in social and applied epistemology and the philosophy of cognitive science. His recent publications include “Real Fakes: The Epistemology of Online Misinformation,” “Epistemic Domination,” and “Beyond Belief: On Disinformation and Manipulation.”
Routledge Studies in Epistemology

Illuminating Errors
New Essays on Knowledge from Non-Knowledge
Edited by Rodrigo Borges and Ian Schnee

Digital Knowledge
A Philosophical Investigation
J. Adam Carter

Seemings and the Foundations of Justification
A Defense of Phenomenal Conservatism
Blake McAllister

Trust Responsibly
Non-Evidential Virtue Epistemology
Jakob Ohlhorst

Rationality in Context
Unstable Virtues in an Uncertain World
Steven Bland

Seemings
New Arguments, New Angles
Edited by Kevin McCain, Scott Stapleford, and Matthias Steup

The Epistemic Injustice of Genocide Denialism
Melanie Altanian

Misinformation, Content Moderation, and Epistemology
Protecting Knowledge
Keith Raymond Harris

For more information about this series, please visit: https://www.routledge.com/Routledge-Studies-in-Epistemology/book-series/RSIE
Misinformation, Content Moderation, and Epistemology
Protecting Knowledge

Keith Raymond Harris
To my parents and to Anna
Contents

Acknowledgments viii

Introduction: Knowledge under threat 1

PART I
The threat of misinformation 15

1 How misinformation prevents knowing 17
2 Conspiracy theories and runaway skepticism 40
3 Ambiguity, fakery, and social evidence 61

PART II
The promise of content moderation 81

4 Damned if you do: The case against content moderation 83
5 Damned if you don’t: The case for content moderation 105
6 Collaborative content moderation 126
   Epilogue: Online, digital soldiers 147

Index 154
Many people and institutions helped to make this book possible. I would first like to thank Tobias Schlicht, who gave me the opportunity to write this book and whose seminar on social media sparked many of the ideas developed here. I would also like to thank participants in several conferences—including the 2022 Political Epistemology Network, the Visual Trust conference on Image-making, Changing Minds Online, and the Rijeka conference on Hate Speech, Fake News, and Freedom of Speech. Their comments on earlier material helped to give shape to this book.

I would like to thank Andrew Weckenmann for his early interest in and support for the project and Rosaleah Stammler for her support throughout the publication process. I would also like to thank four anonymous referees for their feedback on early material for this book, which helped to shape the direction of the text. Thanks also to a further anonymous referee, whose extensive comments on an earlier draft of the manuscript proved a substantial help.

I owe a special thanks to the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen for funding the INTERACT! project (PROFILNRW-2020-135) at Ruhr-Universität Bochum, through which this project, including its open access content, was funded.

Last, but not least, thanks to my dog, Buffy, who was a constant presence and an occasional welcome distraction throughout the writing process.
Introduction

Knowledge under threat

Some time ago, I returned to the United States to attend a friend’s wedding. Early in my pre-wedding haircut, I made a mistake with which many philosophers will be familiar: I mentioned that I was a philosopher. Worse still, I mentioned that I was a philosopher whose work dealt, in part, with conspiracy theories. This stopped my hairdresser mid-snip. “Have you heard of NESARA?” she asked eagerly. NESARA, short for National Economic Security and Recovery Act, refers to a supposed plan to radically overhaul the US financial system and is highly anticipated in certain fringe corners of the internet. Conversational journeys that start with NESARA don’t typically circle back to the safer shores of typical small talk and, over the course of my haircut, I was apprised of theories about government conspiracies and hidden extra-terrestrials, coupled with the insistence that “It’s not about red versus blue, it’s about light versus dark.” Because it is not advisable to argue from the hairdresser’s chair, I kept my protestations to the gentle suggestion that “It’s good to keep an open mind.”

Many readers have probably experienced similar interactions. Indeed, for many of us, run-ins with extraordinary beliefs in both our personal lives and in political contexts appear to be increasingly common. It is widely thought that such beliefs reflect the influence of misinformation, and especially the prevalence of outlandish misinformation on social media. Especially because some such beliefs appear to be dangerous, it is often thought that there is cause for measures—including the use of social media content moderation—to suppress the spread of misinformation. But the sensational nature of such beliefs, which often attract an almost voyeuristic fascination, as well as dramatic events that—like assault on the US Capitol in January 2021—appear to be inspired by such beliefs, can distract from some of the more subtle effects of misinformation. In this book, I argue that, without a more complete understanding of the effects of misinformation, attempts to deal with the challenge may exacerbate its ill-effects. In particular, I argue that heavy-handed attempts to suppress misinformation through content moderation run the risk of amplifying some of its worst effects.
2 Introduction

This book is a study in epistemology, that is, the theory of knowledge. I thus focus on the effects of misinformation and content moderation on the attainment of knowledge. The aim of this book is to explore how misinformation threatens knowledge and how content moderation can either mitigate or worsen this threat. Ultimately, I argue that careful attention to knowledge and the conditions for its attainment can help to inform approaches to content moderation that extend, rather than supplant, the agency of social media users.

I.1 The promise of social media and the threats of misinformation

Social media is commonly defined in terms of its user-driven nature (Fuchs, 2014; Kaplan & Haenlein, 2010, p. 61). In contrast to legacy media, which is characterized by one-directional messaging from the few to the many, social media allows ordinary people to generate and broadcast their own content to large audiences. For many people, social media is creatively empowering. Twitter users can write and share jokes that go viral and come to the attention of millions of other users. Amateur photographers and other artists can use Instagram to share their work. TikTok users can record videos that help them to share their tastes and creativity.

Social media is not just creatively empowering, but epistemically empowering. Aided by other technologies, especially smartphones and their built-in cameras, social media allows for the rapid distribution of information. Thus, for example, social media has been instrumental in the distribution of evidence of police misconduct and has thereby shaped subsequent protest movements and calls for greater accountability. Generalizing beyond such concrete cases, the epistemic power of social media consists, in part, in the enhanced ability of ordinary individuals to easily broadcast information to recipients across the planet. Social media platforms thereby go beyond merely connecting people. By allowing for the rapid spread of information, such platforms allow individuals to open windows into their own worlds, through which countless others can look.

But the social dimensions of social media go further still. Social media provides not only for the sharing of experience, but for the shared experience of sharing experiences. When we encounter jokes on Twitter, we thereby connect not only to the joke-teller, but also to the many other people who encounter the same joke. Often, as one watches a livestream on social media, one simultaneously sees the live commentary and emoji reactions of other users. Social media platforms, then, are both conduits for information and shared theaters in which to consume it.

As I will explore throughout Part I, this dual nature of social media platforms gives rise to many epistemic challenges that exist alongside the benefits highlighted thus far. We will see, for example, that the visibility of others’
reactions has, sometimes for better and sometimes for worse, the potential to guide our interpretations of content accessed on social media. More immediately, social media allows not only for the rapid spread of creative products and valuable information, but also for the spread of misinformation.

That the spread of misinformation on social media represents a significant threat to individual and collective knowledge, to decision-making, and even to democracy is a point largely taken for granted in some quarters. Thus, the challenge of misinformation and how best to confront it has been widely discussed among journalists, academics, and politicians, especially since 2016—a year which saw both Brexit and the election of Donald Trump to US President. However, it is worth acknowledging from the outset that the epistemological and political significance of misinformation has recently faced significant scrutiny. Many academics have argued that concerns about the spread and influence of misinformation have the character of a “moral panic” (Altay et al., 2023; Carlson, 2020; Jungherr & Schroeder, 2021) in which the actual impacts of misinformation are blown out of proportion. Some key data in support of this revisionary response include the observations that misinformation is principally consumed (Grinberg et al., 2019) and shared (Osmundsen et al., 2021, p. 1005) by a small minority of social media users.

Although the revisionary response is a useful corrective to sometimes exaggerated claims as to the severity of the misinformation problem, there are several reasons to think this response is somewhat too optimistic. First, studies of the prevalence of misinformation tend to focus on extreme cases. For example, studies of the spread of fake news and other forms of misinformation tend to use relatively narrow definitions of the term such that only content from consistently unreliable publishers counts as fake news. Second, many studies of misinformation focus narrowly on fake news and may underappreciate the impacts of other forms of misinformation. Third, even if few people are deceived by misinformation, small numbers may be enough to have major impacts in some cases, including in the contexts of competitive elections (van der Linden, 2023). Fourth, studies of the impacts of misinformation tend to focus on the western context, but these impacts may be especially potent in geopolitical contexts in which media ecosystems are comparatively underdeveloped. Fifth, even if the role of misinformation in driving large-scale political events is sometimes overstated, misinformation seems to have important effects on individuals’ private lives. A subreddit titled r/QAnonCasualties, for example, details stories of estrangement and the destruction of personal relationships due to loved ones’ devotion to the QAnon conspiracy theory. Such private costs are difficult to study, but of great significance to the people that experience them. Finally, it is consistent with the concrete impacts of misinformation being limited that the specter of misinformation reduces individuals’ trust in online information, generating subtle or extreme doubts that compromise
their abilities to enjoy the full epistemic benefits of social media. This last point, it should be noted, cuts in multiple directions. Both misinformation and exaggerated reports as to the prevalence and impact of misinformation can drive a loss of trust.

In fact, I argue in what follows that reduced trust is one of three prongs by which misinformation threatens knowledge. In the next three sections, I briefly introduce the three-pronged threat of misinformation by way of example and analogy.

I.2 The gunman

A criminal affidavit reveals that, on December 1, 2016, Edgar Maddison Welch received the following text message:

Tell me we r going to save the Indians from the pipeline.

Welch responded with the following pair of messages:

Way more important, much higher stakes
Pizzagate.

As the surrounding string of messages indicates, Welch had been convinced by online misinformation of the need to take drastic action. Whereas Welch’s text exchanges suggest some pre-existing interest in protesting against the Dakota Access Pipeline, this interest was quickly superseded by a plan that would land Welch in police custody and would bring international attention to the false and outlandish Pizzagate conspiracy theory.

The substance of the Pizzagate conspiracy theory is that high-profile Democratic politicians and operatives used a range of locations, including the Comet Ping Pong pizzeria in Washington, D.C., for sex trafficking children. Welch arrived, armed, at Comet Ping Pong just three days after participating in the text exchange above. As he drove to the pizzeria, he recorded messages for his family. These recordings included the following message to his children:

Like I always told you we have a duty to protect people who can’t protect themselves…I hope you understand that one day.

Welch’s conduct reflects a sincere belief in the Pizzagate conspiracy theory—a belief developed through consumption of misinformation on YouTube and other online platforms.

While Welch’s incursion into Comet Ping Pong fortunately ended without injury to himself or any patrons or employees, the case illustrates how
quickly misinformation can deceive its consumers, thereby motivating reckless, violent, or otherwise counter-normative behavior. Welch was not alone in being misled by the Pizzagate allegations. Believers in Pizzagate, and in the QAnon conspiracy theory that grew out of it, have harassed restaurant employees (Rosenberg, 2016), committed murders (Vigdor, 2021; Watkins, 2019), and stormed the US Capitol (Rubin et al., 2021) in the name of these theories. Unsurprisingly, it is the tendency of misinformation to cause false beliefs and thus reckless action that has received the lion’s share of attention. However, the threat of misinformation cannot be understood solely in terms of causing false beliefs.

I.3 The holdout

In 1987, the Sydney Morning Herald reported the death of Norio Suzuki as follows:

Yeti Hunter Dies.

Suzuki had indeed died in an avalanche while searching for a Yeti in the Himalayan mountains. This search for the mythical creature might appear less quixotic when one considers that Suzuki had previously managed to find a target nearly as elusive as the Yeti: Second Lieutenant Hiroo Onoda.

Onoda was an intelligence officer for the Japanese Imperial army during the Second World War. As part of his mission, Onoda was sent to Lubang Island, located in the Philippines, in 1944. Suzuki found him on this same island, roughly thirty years later. To that point, Onoda remained committed to his original mission, unaware that the war had long since ended. Only after the Japanese government sent Onoda’s former commanding officer to officially relieve him of duty did Onoda accept that the war was over.

Prior to Suzuki’s discovery, many unsuccessful attempts had been made to convince Onoda of the war’s end. These are detailed in Onoda’s (1999) autobiography No Surrender: My Thirty Years War in a chapter entitled “Faked Messages.” Signed photographs of Onoda’s family did not convince him, nor did newspapers illustrating the post-war peace that had settled over Japan. In each case, Onoda suspected that the supposed proof was somehow faked. Onoda attributed his suspicions to his intelligence training in the Futamata class at the Nakano military intelligence school:

I had been taught at Futamata always to be on the lookout for faked messages, and it did not seem to me that my attitude was overly cautious...I still remembered learning at Futamata about a fake message that had made it easier for Germany to overrun France in 1940.
6  Introduction

Onoda’s training taught him not to take anything at face value, and instead to consider how various communications might be moves in the game of intelligence and counterintelligence. Even seeing his own brother on the island with a search party was not enough to overcome Onoda’s suspicions:

A man was standing on the top of Six Hundred speaking earnestly into a microphone. I approached a point about a hundred and fifty yards away from him. I did not dare go nearer, because I would have made too good a target.

I could not see the man’s face, but he was built like my brother, and his voice was identical.

“That’s really something,” I thought. “They’ve found a Nisei or a prisoner who looks at a distance like my brother, and he’s learned to imitate my brother’s voice perfectly.”

The man started to sing, “East wind blowing in the sky over the capital...” This was a well-known students’ song at the Tokyo First High School, which my brother had attended, and I knew he liked it. It started out as a fine performance, and I listened with interest. But gradually the voice grew strained and higher, and at the end it was completely off tune.

I laughed to myself. The impersonator had not been able to keep it up, and his own voice had come through in the end.

Only later did Onoda learn that his brother’s voice had cracked due to overwhelming emotion.

Onoda’s imperviousness to evidence of the war’s end was costly. He describes this realization as follows:

For thirty years on Lubang I had polished my rifle every day. For what? For thirty years I had thought I was doing something for my country, but now it looked as though I had just caused a lot of people a lot of trouble.

Onoda poignantly describes the source of his confusion and that of another soldier that was his companion for many years:

Kozuka and I had developed so many fixed ideas that we were unable to understand anything that did not conform with them. If there was anything that did not fit in with them we interpreted it to mean what we wanted it to mean.
Onoda was primed for suspicion by a history of education in intelligence, a domain in which fakery is to be expected. A formal education in intelligence is atypical of the general population. However, widespread awareness of the existence of fake news and manipulated media can likewise prime ordinary people for suspicion, especially of information that conflicts with their existing beliefs. Similarly, while most of us need not worry about fake brothers, we might worry about fake online persons in the form of bots and trolls. Such awareness might arise through direct experience. Noticing fake news in one’s social media feed, for example, one might come to fear that there is more fake news present that one has not yet recognized as such. Consequently, one might come to doubt the reliability of even legitimate news.

Especially since the political events of 2016—most notably Brexit and the election to President of Donald Trump—the popular press has been heavy with stories about fake news, bots, Russian trolls, deepfakes, and other forms and disseminators of misinformation. Such stories no doubt inspire some to be more vigilant in their information consumption and thereby help to avoid deception. However, such vigilance can easily slip into excessive skepticism. Focusing on the possibility of deception might lead audiences to make similar mistakes to Onoda’s, that is, to reject even accurate information. In this way, even accurate, well-intentioned investigations of fake news and related phenomena can contribute to skepticism toward legitimate sources of information. It is thus essential not to regard the threat of misinformation solely in terms of its deceptive potential. It is likewise essential not to exaggerate the deceptive potential of misinformation, thereby fueling excessive skepticism.

I.4 The counterfeiters

In September of 1939, the Nazi government began to devise a plot to overcome one of its most formidable foes: The British economy. The twentieth century had already seen the industrialization of warfare manifest in the mass production of munitions. Arthur Nebe—head of Germany’s criminal police—proposed an innovative application of German industrial capacity. The plan was to counterfeit and distribute British banknotes on a massive scale, thereby thrusting devastating inflation onto the British economy. Although the plan was never fully implemented, the attempt to mass produce convincing British banknotes was ultimately successful. The plotters later turned their attention to the forging of American banknotes.

Several of those involved expressed concerns about the plot, either privately or to the other plotters. Two of the most serious objections concerned the potential for it to backfire. In his diary, Joseph Goebbels wondered what
would happen if the British used the same tactic in retaliation (Malkin, 2006, Chapter 1). Others worried that discovery of the plan would leave the Germans with reputations as counterfeiters, thus doing long-term damage to the value of German currency.

At the heart of the Nazi plot, as well as the potential objections to it, is the recognition that the proliferation of fake currency degrades the value of real currency. Something similar is true of evidence. Ordinarily, news reports provide good evidence in favor of the claims reported. Ordinarily, photos and videos provide good evidence that the events they depict really occurred. But, where fake news, fake photos, and fake videos abound, even their authentic counterparts lose much of their evidential value. Misinformation, in effect, is counterfeit information (Fallis & Mathiesen, 2019) and tends to have an effect on legitimate evidence that is analogous to the effect of counterfeit currency on authentic currency.

Suppose that one forms a true belief that the President of the United States committed a verbal gaffe based on video footage that seems to show this. Suppose, further, that the video footage is authentic and has not been edited in any misleading way. Even in this case, one arguably does not know that the President committed a verbal gaffe. After all, it is plausible enough that the President’s political foes might have concocted fake video footage or edited authentic footage to give a misleading impression. Even this possibility—increasingly realistic in light of novel techniques for cheaply and quickly editing video footage—is arguably enough to prevent one from having knowledge. This is because a given observation serves as good evidence for a given hypothesis only to the extent that the observation would be relatively unlikely if that hypothesis were false. For this reason, the realistic possibility that there is inauthentic evidence reduces the value of even authentic evidence in a way that undermines knowledge acquisition.

That misinformation reduces the significance of evidence is not a novel point. It is a familiar lesson that crying wolf in the absence of wolves reduces the evidential value of one’s cries. It is not merely that falsely crying wolf encourages skepticism of even true reports in the future, such skepticism is rational in light of the reduced evidential value of one’s cries. But counterfeits—whether informational or monetary—do not simply reduce the value of the authentic counterparts issued by the same authors. Fake British banknotes would reduce the value of even legitimately issued banknotes. Likewise, if they cannot be distinguished, fake news issued by unscrupulous parties can degrade the evidential value of news reports issued by scrupulous parties.

Suppose that the Nazi plot had been carried to fruition and that massive quantities of realistic counterfeit British currency had been dropped over British cities. It would then be reasonable for British citizens to apply extra
scrutiny to payments before accepting them and perhaps to avoid accepting them altogether. Likewise, it is often reasonable for those aware of the possibility of misinformation to apply extra scrutiny to information retrieved online and perhaps to avoid forming beliefs based on this information. Misinformation gives rise to skepticism not because it causes people to be irrational, but because it makes skepticism rational.

I.5 Content moderation and its discontents

Governments take counterfeiting very seriously. Those that engage in counterfeiting and indeed those that merely pass on counterfeit banknotes are typically subject to serious penalty, and counterfeit banknotes themselves are typically destroyed. Such heavy-handed measures are not the only means by which the destructive impacts of counterfeit currencies might be mitigated. In principle, ordinary citizens might be trained in the detection of counterfeits, or equipped with devices for distinguishing between real and fake currency. However, such alternatives would ask much of ordinary people and would slow down many financial transactions. In contrast, the strategy of deterrence and destruction functions to protect the integrity of currency quite generally, without demanding excessive vigilance on the part of ordinary persons.

Given the analogy between counterfeit currency and misinformation, it is at least initially plausible that a comparably heavy-handed approach would be effective and appropriate for mitigating the threats posed by misinformation. If it is appropriate to destroy counterfeit banknotes to protect the value of currency, then it is likewise appropriate to protect knowledge by destroying misinformation—or at least by removing misinformation from the social media platforms on which it proliferates. In this way, content moderation—in the form of interference with who and what can appear on social media and in what context—appears to be an appropriate response to the challenges posed by misinformation. This is not to say that more gentle responses, involving for example better education in information literacy, are unnecessary or undesirable. It is instead to say that content moderation is at least one valuable weapon among the arsenal of those that can be deployed to combat misinformation’s ill-effects.

Yet the use of content moderation against misinformation has proven highly controversial. Some objections to content moderation appeal to the inalienable right to free expression. Other objections are more specifically epistemological, alleging for instance that content moderation is, or could be, used to suppress the truth. It is such epistemological issues concerning the threats of misinformation and the efficacy of content moderation in mitigating these threats with which this book is concerned.
Here is the plan for the book. In Part I, I develop an account of how misinformation threatens knowledge, and how the unique social feedback encountered on social media exacerbates this threat. This begins, in Chapter 1, with a distinction between three threats posed by misinformation: the deceptive threat, the skeptical threat, and the epistemic threat. I then argue that misinformation’s threat to knowledge does not, strictly speaking, depend on the real existence of misinformation. Even the mere propensity of misinformation to exist and indeed mere concerns as to the possible existence of misinformation are enough to threaten knowledge. Chapter 1 concludes with a discussion of some limitations on misinformation’s threats to knowledge.

In Chapter 2, I discuss the skeptical threat of misinformation at greater length, with a special focus on the role of conspiracy theories in driving skepticism of official explanations and evidence presented by epistemic authorities more generally. I argue that the ill-effects of conspiracy theories do not depend on the irrationality of individual believers, but rather that conspiracy theories can make rational people believe absurd things and, often as importantly, fail to believe well-evidenced claims.

Chapter 3 concerns the role of social evidence in the discrimination between accurate and inaccurate information. Such evidence takes many forms on social media, including testimony, likes, and shares. I argue that, while social evidence can in principle help to distinguish between accurate and inaccurate information, the significance of this evidence is compromised by several factors. These include the ambiguity of certain forms of social evidence, the varied motivations that individuals have for engaging in social media communication, and the distorting influence of trolls and bots. Finally, I argue that epistemic virtue on the part of individuals is not enough to secure the value of social evidence.

In Part II, I turn to the effectiveness of content moderation as a response to the challenges of misinformation and to positive proposals as to how content moderation can best be put to this end. This begins in Chapter 4, where I develop several epistemic arguments against the labeling and removal of misinformation from social media. These arguments highlight the likelihood that content moderation will fail to be comprehensive and the lack of trust that many individuals have toward content moderators. I argue that these factors not only limit the effectiveness of content moderation as a way of mitigating the threats of misinformation, but that, in light of such factors, content moderation can exacerbate the threat to knowledge. Finally, I argue that content moderation threatens to reduce the evidential value of testimony and apparent consensus insofar as content moderation amounts to tampering with the social evidence.
Then, in Chapter 5, I argue that there is good reason to engage in content moderation, despite the concerns raised in the previous chapter. In part, this is because social media platforms must exert control over the spread of content, and thus—even absent explicit attempts to mitigate the spread of misinformation—some of the epistemic ill-effects of content moderation are unavoidable. Furthermore, content moderation can, at least in many cases, mitigate the three-pronged threat of misinformation. Additionally, individuals are subject to a range of limitations that promote the spread and consumption of misinformation and that can be reduced through content moderation. Finally, the concern that content moderation amounts to social evidence tampering is of limited importance, as individuals already struggle to assess the weight of such evidence.

Chapter 6 offers several proposals as to how content moderation can be conducted to combat the three-pronged threat of misinformation while minimizing the epistemic ill-effects of content moderation itself. Generally, I argue that content moderation ought to aim at enhancing the epistemic agency of ordinary individuals as both consumers and moderators of content. This can be accomplished, in part, by enhancing the quality of social evidence and by assisting individuals in assessing the track records of various sources.

In the epilogue, I emphasize that misinformation is not simply an epistemological problem. Very often, the consumption and sharing of misinformation reflects underlying individual, social, and political problems. Such problems must be confronted but, as I argue, protecting knowledge is an essential part of doing so.

I.7 Epistemology, social media, and content moderation

Social media misinformation and content moderation may strike some readers as strange topics to address through a philosophical lens. Often, philosophy is associated with perennial issues like the meaning of life and the nature of fundamental concepts like knowledge and justice, rather than issues introduced by new technologies. Yet, if philosophy is to be relevant, it must engage with changing conditions in the world, including novel technologies. Moreover, both misinformation and content moderation are natural topics for epistemology. Much of epistemology is devoted to exploring ways in which appearances might systematically diverge from reality and the consequences of such possibilities for human knowledge. Thus, for example, the ancient skeptics considered how the peculiarities of our sense organs and the conditions of our bodies might distort our perceptions of the world. The ancient Chinese philosopher Zhuangzi and later the early modern philosopher René Descartes considered the possibility that our experiences of the world were mere dreams or perhaps due to the
machinations of a powerful and deceptive demon. Twentieth-century philosophers modernized such anxieties, considering the possibility that our experiences might be manufactured by deceptive neuroscientists. More recently, epistemologists have devised elaborate thought experiments to consider how fakes in one’s environment might compromise the acquisition of knowledge. Epistemological consideration of social media misinformation is thus not a radical departure from the discipline’s history. Instead, it amounts to the application of existing conceptual tools to situations encountered in the real world.

Content moderation is likewise a suitable object of epistemological theorizing, albeit one that has thus far received limited attention from epistemologists. The epistemology of content moderation is not wholly uncharted territory, as it is partly an extension of long-running discussions of the epistemology of free speech and censorship. Philosophers have argued that free speech carries important epistemic benefits and that the restriction of speech does great damage to the individual and collective pursuit of knowledge (Wright, 2021). Social media offers ordinary individuals radically expanded abilities to broadcast their ideas, while content moderation potentially restricts such abilities. In short, while social media creates new opportunities for freedom of expression, it also affords new means of control. Exploring these issues through an epistemological lens serves to bring theoretical tools to bear on issues of practical significance while also promising to enrich epistemology itself.

All this said, social media and content moderation present unique challenges for epistemological study, insofar as certain relevant facts—concerning the existence, popularity, and policies of various platforms—are subject to rapid change. Thus, although I refer to specific platforms and policies where appropriate, this book does not attempt to provide an in-depth study of the specificities of content moderation policies on various platforms. Rather, I aim to assess, through an epistemological lens, various types of content moderation policies that have been used and that might be used in the future.

Notes
1 Relatedly, there is evidence that warnings about misinformation (Van Der Meer et al., 2023) as well as attempts to train individuals to better identify misinformation (Modirrousta-Galian & Higham, 2023) reduce credulity toward misinformation at the cost of also reducing credulity toward legitimate information.
2 But see Karen Frost-Arnold (2023) for a rare exception.
3 For example, the name of Twitter was recently changed to “X.” Because relevant theoretical and empirical studies referenced throughout this book were carried out when the platform was called Twitter, I continue to use that name to refer to the platform.
References


Introduction


Introduction


How misinformation prevents knowing


Conspiracy theories and runaway skepticism


Ambiguity, fakery, and social evidence


---

**Damned if you do**


Damned if you don't


Collaborative content moderation


Epilogue

Whitehurst, L. (2022, December 16). QAnon Follower Who Chased Capitol Officer on Jan. 6 gets 5 years. PBS NewsHour. https://www.pbs.org/newshour/politics/qanon-follower-who-chased-capitol-officer-on-jan-6-gets-5-years