Fighting falsity:

Essays on deceiving, objecting, and conspiring

Rory Aird
MSc, M.A.

A thesis presented for the degree of
Doctor of Philosophy

School of Humanities
College of Arts
University of Glasgow
August 19, 2025

Abstract

This thesis is comprised of five discrete but related chapters in applied, social, and political epistemology. Specifically, most of the thesis is centred around fighting falsity; that is, for example, what should you do when one someone asserts something wrong, outlandish, or flat-out dangerous to you or those around you? What is the best course of action if you come across such a claim on the internet? What sort of claims even are wrong, outlandish, or flat-out dangerous? I open with a discussion of bullshit and identify a new phenomenon I call "hedged bullshitting" which I argue is a superlative form of deception we ought to be concerned about. I turn in chapter two to a discussion of what makes for an epistemically good objection (and forbearance from objecting), employing the influential performance-normative framework from virtue epistemology to answer this. I then call for caution in chapter three; objecting can and will go wrong, especially when it comes to certain controversial false assertions, thus we ought to not be so laissez-faire when it comes to engaging with interlocutors. In light of this call for reticence, I explore a different method of fighting falsity in chapter four: censorship and no-platforming. While potentially effective strategies, I argue they make for hiltless swords—they cannot be safely used. In the final chapter, I offer a broad critique of the philosophy of conspiracy theories, rejecting almost every widely held assumption in the literature. I offer my own, more satisfactory definition of conspiracy theories that I argue will best advance the literature.

Contents

List of Tables

Acknowledgements

They say it takes a village to raise a child. I wouldn't know. But, to throw my own hat in a syntactically similar proverb ring, I would contend that it takes a university to write a PhD thesis. So, perhaps strangely, my first port of call in this acknowledgements is not a person or a people, but a place. For me and my PhD thesis (and my Master's and my undergrad) that place was the University of Glasgow. Glasgow Uni—and the city itself—have been my home now for nearly a decade and while it would likely be a slight exaggeration to say I have loved *every* minute of it, I can confidently say at this point that I still would not change a single second. Thank you, Glasgow, for having me and letting me do my PhD in the best city in the world.[1] I hope one day to get the chance to come back and stay again, for keeps this time. Nevertheless, as Glasgow City Council likes to remind us, *"People Make Glasgow"*, so it is to those people that made my time at Glasgow we now turn.

At the start of my Master's degree, I decided to set my sights on doing a PhD in philosophy. The idea had been kicking around my head for a while, essentially since being a first or second year undergrad, but I don't know if I really believed I would do it, or if it was just a random pipe dream. To be honest, I didn't really know what a PhD even was.

Luckily for me, on the 27th of October 2020, despite not knowing me from, well, anyone, a certain Adam Carter answered an email with the subject "Epistemology, PhDs, and other things" from a completely clueless Master's student. His reply opened, "Hi Rory, happy to help!" I had no idea how true that was. From that moment on, he did help and his help permeates every single part of this thesis. From right at the start where he roped in Emma Gordon to be my other supervisor and they both helped me revise a reasonable PhD proposal from the, let's say, rough outline I had sent them, to every round of comments on every paper and R&R and job proposal and cover letter, their help has been beyond invaluable—and never longer than a week in response! I owe them an unrepayable debt for all their kindness, support, enthusiasm, and everything else. Perhaps the best compliment I can afford them is that if I am ever lucky enough to supervise a PhD student, and I do half as good a job as they did with me, I can be assured that I am an excellent supervisor, one of the

1 It's my acknowledgements—I can claim what I want.

absolute best.

Around the same time, I had my first encounter with another person who would go on to play a key role in this thesis. I refer, of course, to Ross Patrizio. There is scarcely a sentence in every chapter to come that has not in some way or another benefitted either directly or indirectly from conversation, comments, or collaboration with him. Going into the office and chatting philosophy for hours (possibly to the detriment of everyone else there) has been a bona fide highlight of the whole experience that I will perennially cherish. We also went out drinks maybe one or two times which was fun. There is no doubt in my mind whatsoever that every piece of work in this thesis would be substantially worse if not for his myriad of contributions. Getting down into the bare-bones legalities of the matter, I owe him.

The acknowledgements so far have been fairly effable. The next thanks, to Ali McKinlay, may be admittedly more ineffable, but undoubtedly no less valuable. Thank you for being my best friend, for keeping me grounded, and for genuinely being the funniest person I have ever met in my life. You played a role in all this I doubt you'll ever know. Hopefully, you don't ever read these words because I'll never live it down.

Exactly where to turn to next is difficult because there are just so many important people to thank. Maybe I'll just try winging it. To all my wonderful friends in the philosophy department, thank you for all the amazing times. I've always said that the Friday night post-seminar pub helped me to maintain my sanity throughout the PhD, and I do think that was true. But, to be honest, what value would such sanity have had without those nights out anyway? To everyone over the years—Annalisa, Calum, Carina, Fin, Giorgia, Graham, Ísak, Lisa, Louis, Martin, Matthew, Pat, Penelope, Rocco, Will—thank you for putting up with me (and Billy Joel).

To all the other amazing staff in the department past and present (who are also my friends but I'll separate for convenience), I owe so much to so many of you. Everyone at team COGITO has made doing the PhD such a pleasure; Mona Simion and Chris Kelp for all their support and help with jobs and putting on workshops and coming along to things we organised; all the fantastic postdocs, Oscar, Tim, CWK, Giada, Josh, Lilith, Matt, and everyone else throughout the years who read my papers and gave me time when they clearly had enough on their plates. Further thanks to Ben Colburn for being a genuinely good person who is always willing to help out; Robert Cowan for putting on such an excellent opening lecture to Philosophy 1K that a 1st year student who was taking philosophy as a third subject decided to change their whole degree; Neil McDonnell for being there at some crucial times right at the beginning; Joe Slater for being a moral exemplar and a great friend; I could go on for (roughly) forever. Suffice to say, I will never be able to adequately thank everyone, but I have tried.

We are nearly done. I mentioned earlier that I also did my undergraduate and Master's degrees at Glasgow Uni and so have lived here for nine years. One incred-

ible benefit of that is that I have been able to easily maintain friendships in Glasgow outside the philosophy department. To Iain, for the laugh; Johnny, for keeping me on my toes, no matter what time it was; Pete, for luminiferous aether, among other things; Tom, for all the food recommendations. Everything was a lot easier knowing some ethically sourced ones were just round the corner.

The penultimate spot goes to my mum and dad. Thank you for all your support—it can't have been easy my doing something often impenetrable to those already within it.

There's only one person left to thank. And that is you, dear reader, for reading. I didn't write this all this *just* for me—even if mainly so. Enjoy!

Author's Declaration

I declare that all the work in this dissertation is the result of my own efforts, except where properly indicated by means of quotations and references. This work has not been submitted for any other degree of professional qualification. This work has received generous and invaluable financial support from the Scottish Graduate School for Arts and Humanities and the Royal Institute of Philosophy.

Statement of Length

This dissertation, including footnotes and references, contains around 73,500 words. It meets the minimum word limit set by the University of Glasgow College of Arts (70,000) and does not exceed the maximum word limit (100,000).

Published Material

At the time of submission of this thesis, Chapters 3 and 4 have been accepted for publication or published in the following journals:

- Aird, Rory. 2023. "A puzzle of epistemic paternalism." *Philosophical Psychology* 36 (5):1011–1029
- Aird, Rory. Forthcoming. "On the perils of engaging." *Episteme*

Introduction

> Proof is boring. Proof is tiresome. Proof is an irrelevance. People would far rather be handed an easy lie than search for a difficult truth, especially if it suits their own purposes.
>
> Inquisitor Glokta, *The Last Argument of Kings*

Upon meeting new people, they will often ask me what I do for work. "I'm actually a PhD student," I usually reply, almost apologetically.

"Oh really?" eyebrows raised, they might respond, "What in?"

Already, I am somewhat hesitant—"philosophy PhD student" potentially carries with it some unflattering assumptions. So, it is with a not insignificant amount of trepidation that I answer: "…Philosophy."

At this point, in my experience, there are three options: 1) the person loses interest in response to this answer and the conversation ends or moves elsewhere; 2) they are genuinely curious and inquire further about the specific area of philosophy; or 3) disaster strikes and I find myself being quizzed on Camus or Nietzsche. When the latter happens I am usually quickly outed as a fraud—"what sort of *philosophy* PhD student wouldn't know anything about Derrida?" they might exclaim. At least it's over quickly.

Nevertheless, let's take a closer look at 2). "Well…" I begin to explain, instinctually hedging tonally, "The *general* area is epistemology…"—I notice eyes glazing over—"*But*… I actually do quite applied stuff: about people who believe and say seemingly irrational and outlandish and wrong things and what we should do when confronted with such ideas and how best to tackle them and that sort of thing." I end rather breathlessly.

Perhaps surprisingly (or perhaps unsurprisingly), people are often fairly interested at this point. And this does make some sense: in the past few years, those who believe and say seemingly irrational and outlandish and wrong things have gone from the fringes of forgettable movies to front and centre of our political and social lives. "Fake news" in 2016 may have been the spilling-over candle that started the "post-truth" fire, but the Covid-19 pandemic was the nearby stack of dynamite caught in the flames a few years later. The subject of this thesis is somewhat pre-

dictable really, considering it all began in the throes of the pandemic. I remember seeing everywhere claims about how the restrictions were pointless, the hospitals empty, Covid a hoax. *How can people think this?* I would despair. *Why are they so confident? What should I do? What can I do?* As I was asking myself such questions, so too, I think, was philosophy—and no area more so than *epistemology*. Suddenly, the basic building blocks of epistemology were thrust into the limelight: justification, truth, belief, knowledge—Barack Obama even called it an "epistemological crisis." (Goldberg 2020)

To abuse some creative license, we can imagine that epistemology was previously interested in positive elements—how to attain justification, how to get and pass on knowledge, how to be rational. Suddenly, however, the focus was flipped: how falsehoods can defeat justification, how we can lose knowledge, how people are irrational. And along with it came a plethora real-life examples. We no longer needed to invent Gettier (1963) cases to discuss these issues, you just had to go outside or go on the internet. The incredible apparatus of analytic epistemology was perfectly placed to be fruitfully applied to matters of the zeitgeist. Indeed, in the thesis to come, I will often use real, if sometimes pseudonymous, cases to discuss these issues and try to provide philosophically robust analyses and solutions.

How can people think x false belief? I despaired. *How can they be so confident in y irrationality? What should I do? What can I do?* Perhaps you have also despaired over such questions. This thesis is my attempt in providing an answer to (some of) them. And while it is comprised of five discrete (although obviously related) chapters, I will tease out a narrative between them (in *Linking Interludes I–IV*) and I hope that they can be taken both individually *and* as a coherent whole.

In Chapter 1, I discuss bullshit. More specifically, I discuss a phenomenon I identified that I call *hedged bullshitting*. In short, it is when someone's bullshitting includes terms like "I think" or "possibly"—viz., *hedges*. This has—as will be a common thread throughout most of my thesis—both theoretical and practical implications. For the former, standard theories of bullshit predict the impossibility of hedged bullshitting despite its clear existence; for the latter, such hedging, I argue, is liberally employed by bullshitters in real life to give them deniability, among some other interesting normative effects.

In Chapter 2, I build a picture of the evaluative normativity of objecting—that is, I give an answer to the question of what makes for an epistemically good objection. My solution employs a performance-normative framework (famously used in virtue epistemology) and argues that a good objection is an apt one. I then turn to a novel area: *forbearance* from objecting, what I call "quiescing", and argue that this normative picture has the resources to tell us what makes for good quiesces as well.

In Chapter 3, I call for caution in objecting to certain claims like climate change denial or anti-vaccine sentiment. I argue that objecting can be a dangerous business

and it can go terribly wrong, potentially resulting in widespread deleterious epistemic effects for the epistemic environment—like third-party bystanders losing their knowledge if one's engagement with the climate change sceptic goes poorly.

In Chapter 4, I explore a different method of combatting claims of climate change denial and the like; namely, widespread epistemically paternalistic policies of censorship and removing individuals' platforms. While these show promise in effectively tackling misinformation, I argue that, due to the only institutions actually capable of implementing such policies being powerful corporations and governments, these policies cannot be used without incurring serious risks.

Finally, in Chapter 5, we turn to the elephant in the room: conspiracy theories. Although they scarcely make a mention up until this point in the thesis, I engage in a full-scale exploration of the philosophy of conspiracy theories, arguing that most wide-held assumptions about definitions and desiderata are mistaken, that we ought to reject every popular view, and that my own, *position to know* account is to be preferred if the literature is to progress.

The introduction is almost complete. Before my thesis begins proper, however, I would like to add one more thing. When writing this introduction, and discussing how misinformation and the like seem truly matters of the present, of the zeitgeist, I was reminded of an apocryphal quote from Socrates. It goes:

> The children now love luxury. They have bad manners, contempt for authority; they show disrespect for elders and love chatter in place of exercise.

It is easy to think that we are the first to go through everything and that nothing has ever been so bad as it is now. I reject that hypothesis. And while this thesis might seem pessimistic at times, we have got through things like this before, and I am confident we will again. So, perhaps the place my thesis begins is quite appropriate; Harry G. Frankfurt's (1986) essay "On Bullshit" opens with the line: "One of the most salient features of our culture is that there is so much bullshit." Indeed.

Chapter 1

Hedging, bullshitting, and hedged bullshitting

Standard theories of bullshitting say that bullshit assertions are uncoupled from truth or falsity. Hedged assertions, on the other hand, are (explicitly) connected to the alethic. Therefore, these standard theories predict that hedged bullshitting is impossible. And yet, bullshitters hedge their bullshit all the time. This chapter has two key theses; first, I show that hedged bullshitting is a bona fide phenomenon and therefore we ought to endorse an alternative account of bullshit that can explain it; second, I argue that not only do bullshitters often hedge their bullshit but the normative effects of hedging are uniquely positioned to assuage key reputational and weakness-to-challenge vulnerabilities associated with consummate bullshitting. Thus, I propose, an understanding of hedged bullshitting is of crucial import—it is a superlative form of deception.

1.1 Introduction

In *On Bullshit*, Harry Frankfurt writes:

> Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic. (Frankfurt 2005, 63)

This seems right; the world is filled with those who incessantly give their uninformed takes and it is often natural to label them as bullshitters. Here are some such examples, taken from social media (Twitter/X), by polemical figures who offer their thoughts tens to hundreds of times a day about nigh on every issue possible:[1]

1 I will not flag the authors nor link to the posts as I do not want to give them further attention. Nonetheless, they are real and can likely be found if one so wishes.

1. The prime minister of Slovakia, Robert Fico, had [sic] been shot. This comes days after Slovakia's courageous rejection of the [World Health Organisation's] audacious Pandemic Preparedness Treaty and International Health Regulations. *We can't be certain if these events are connected*. But whoever is steering the WHO clearly views natural sovereignty as an irritant, and human lives as disposable.

2. I just wonder if the truth is being withheld from us. *I don't know the answer to that*.[2]

3. Google's Gemini AI project doesn't want to display images of white men—even historical figures. *I think I may have figured out why*.

4. Google Sky censored a red dragon in the constellation Virgo. This beast is *probably* the beast from Revelation 12.[3]

5. Trump *might* break 100,000 attendees tonight. Absolutely insane.[4]

I take these to be paradigmatic cases of bullshitting, particularly when we consider the financial mechanics of such social media sites where users are paid per click. Controversy (and sheer productivity) sells, thus for the unscrupulous user, speaking knowledgeably is unlikely to be the main incentive when posting—the truth pales in comparison to the dollar.

So, (1)–(5) above are prime examples of bullshitting. That is not all, however; they are also cases of *hedging*. Note the parts in italics: they all are or contain paradigmatic hedge terms.[5] We will get to precise explications of what hedging entails for speakers and their assertions in due course. Nevertheless, for now, it suffices to say that (1)–(5) certainly appear hedged insofar as the speakers are explicitly flagging uncertainty or lack of knowledge of the asserted content through the use of hedge terms. Call such *possible-falsity-signalling* the following:

> ALETHIC CONNECTION
> A hedged assertion is (explicitly) connected to the alethic.[6]

(1)–(5) are not only *bullshit*, then, but they are also *hedged*. Thus, prima facie, we have cases here of *hedged bullshitting*. This stems from a general rule of hedging; note how, for instance, an assertion affixed with a hedge becomes a *hedged assertion*;[7] how a

2 This is part of a longer speech in response to the race riots in the UK over Summer 2024.
3 Attached to this post is a (real) picture of Virgo in 2017 from Google Sky.
4 Attached to this post is a picture of the rally with 10,000 people. It eventually totalled 21,000.
5 To be clear, the originals were not italicised in any way—I have added that here for clarity.
6 Unless you hold that the addition of a hedge means the statement can no longer be an assertion—see Benton and van Elswyk (2020) or van Elswyk (2023) for more. Nothing hangs on this anyway, the *assertion* in ALETHIC CONNECTION can be replaced with *declarative* or *statement* for the same result in this chapter. In any case, I will use "hedged assertion" and "hedged declarative" interchangeably throughout.
7 Modulo note 6.

hedge added to a lie results in a *hedged lie* (Betz-Richman 2022), or how a directive with a hedge appended turns into a *hedged directive* (Fraser 2010). Following this standard then, when you have hedge terms attached to bullshit, as we do in (1)–(5), you have hedged bullshitting. We can in turn derive a general claim from these cases:

> HB THESIS
> One can hedge one's bullshit assertions.[8]

Prima facie, this is not a particularly exciting nor very strong claim—it is merely the bullshitting variant of the hedge rule fielded above insofar as, say, the *directive* version would read: *One can hedge one's directives*. So, what is of interest here? Why have I flagged these cases of hedged bullshitting? As I will soon show, standard theories of bullshitting (Frankfurt 2005; Stokke and Fallis 2017), what I will call the ALETHIC ACCOUNTS, are committed to the following claim:

> ALETHIC UNCOUPLING
> A bullshit assertion is uncoupled from the alethic.

Therefore, when we combine ALETHIC UNCOUPLING with hedging's ALETHIC CONNECTION and the HB THESIS, we get an inconsistency. Trivially, the same assertion cannot have both properties p and $\neg p$ simultaneously—that is, hedged bullshitting cannot synchronically be both *connected to* and *uncoupled from* the alethic. So, the ALETHIC ACCOUNTS entail that hedged bullshitting is impossible, contradicting the HB THESIS. Thus, one of these three claims has to give. I will argue that we ought to jettison ALETHIC UNCOUPLING and thus reject the ALETHIC ACCOUNTS.

Here is how the chapter will transpire: first, I show in §1.2 that the ALETHIC ACCOUNTS are indeed committed to this ALETHIC UNCOUPLING. In §1.3, I give a full exegesis of hedging and its ALETHIC CONNECTION. Next, I consider and discard some objections that look to solve the above inconsistency by rejecting the HB THESIS (§1.4). Thus, I propose in §1.5 that we endorse a competitor account of bullshit that allows for hedged bullshitting. In §1.6, I discuss the finer details of hedged bullshitting itself, and argue that it is an extremely dangerous form of deception due to the normative effects hedging uniquely has on bullshit. In §1.7, I conclude.

1.2 BULLSHITTING

Above, I suggested that standard theories of bullshitting, the ALETHIC ACCOUNTS, are committed to the following claim:

8 Equally, this could presumably be adduced as any of the following: (i) Hedged bullshitting is possible, (ii) One can engage in hedged bullshitting, (iii) $\exists x(Hx \wedge Bx)$, and so on. I take these all to have similar meanings for my purposes here in that all I want to put forward is that hedged bullshitting is something that can be done and does exist.

ALETHIC UNCOUPLING
A bullshit assertion is uncoupled from the alethic.

In this section, I will demonstrate so. First, however, let me say a little more about this ALETHIC UNCOUPLING. My choice of "uncoupled" was deliberate insofar as its use was intended to bring to mind a railway car being detached from the engine at the front. The thought is an analogy: if the engine at the front is the alethic which guides and pulls along the other cars—i.e., our (non-bullshit) assertions—by being coupled with them, then *bullshit* assertions are cars at the back that have become uncoupled. The uncoupled cars can through happenstance still follow the same path as the coupled cars but the engine is no longer compelling them so. In the same way, bullshit assertions *can* certainly be true (even if more often false), it is just that they are in no way *impelled* by the alethic—they can at any point go down a different track from the train of truth and falsity.

Let us now put this into practice by turning first to the classical analysis of bullshit from Frankfurt (2002; 2005). He proposes that bullshitting is characterised by an indifference towards the truth (and indeed falsity) of what is said:

> It is just this lack of connection to a concern with the truth—this indifference to how things really are—that I regard as the essence of bullshit. (Frankfurt 2005, 33–34)

In order to demarcate bullshit from related areas of deception, it is helpful to directly contrast it with them. Consider the following basic account of lying:

LYING DEFINITION
To make a *believed-false* statement to another person with the intention that the other person believe that statement to be true.[9] (Mahon 2016)[my emphasis]

The difference between bullshitting and lying is that lying expressly involves a positive belief that what one is saying is false; bullshitting has no such belief either way about the truth-value of what is asserted—it is wholly removed from such alethic linkage. Gibbons (2023, 3) compares two politicians: one who deliberately shares noxious rumours about their opponent that they know or believe to be false, and one who shares similarly injurious tales without any consideration of their alethic merits. The thought is that the latter is bullshitting while the former lies.

To return to the rail analogy, lying involves deliberately going down the fork of falsity, but in order to successfully do that one must also know (or at least believe) that the other path is the one of truth—as Frankfurt puts it: "The liar is inescapably

9 This simple account will suffice here but there are, of course, many alternative definitions of lying. See, for instance, Fallis (2009) for a definition connecting lying with Grice (1978) and his maxims, or Stokke's (2013a) account which employs Stalnakerian common ground (Stalnaker 1978). For a nice overview of this literature, see also Stokke (2013b).

concerned with truth-values." (Frankfurt 2005, 51) The bullshitter, however, goes down whatever path they may without any guidance or imperative from the alethic engine, "[their] intention is neither to report the truth nor conceal it." (55) They are "not constrained by the [truth]" (52); they are *alethically uncoupled*.

Let us now briefly look at what I take to be the other main player in the bullshitting literature: that of Stokke and Fallis (2017).[10] As will soon become clear, the bullshit literature post-Frankfurt mostly builds on this alethic-centred core, so ALETHIC UNCOUPLING is maintained, and thus a prediction of hedged bullshitting's impossibility will still be generated.[11]

Stokke and Fallis characterise bullshitting as a kind of indifference towards *inquiry*: "bullshitting [is] a mode of speech marked by an indifference towards contributing true or false answers to [questions under discussion]." (279) The precise account is as follows:

> *A* is bullshitting relative to a QUD q if and only if *A* contributes p as an answer to q and *A* is not concerned that p be an answer to q that her evidence suggests is true or that p be an answer to q that her evidence suggests is false. (279)

Prima facie, it can be difficult to see exactly how this comes apart from the Frankfurtian wholesale alethic separation. In short, Stokke and Fallis' account was built to overcome specific counterexamples where someone might not care about the truth *broadly* but does take care to say only true things in some context (usually for instrumental reasons).[12] Examples of this structure will not prove germane in this chapter, however, so we can mainly set aside the finer details involving the specific answers to QUDs and so on. What is essential to note here is that Stokke and Fallis themselves view their theory as a "conservative extension of Frankfurt's original account," (279) which "preserves the central insight of Frankfurt's influential analysis of bullshitting." (277) Therefore, this theory is mostly a dyed in the wool scion of Frankfurt's,

10 For an alternative (but still a member of the ALETHIC ACCOUNTS) see Moberger (2020). Gjelsvik (2018) has an account of bullshit characterised by a detachment from the institution of the KNOWLEDGE NORM OF ASSERTION which one might think is somewhat different to the ALETHIC ACCOUNTS fielded here. Granted, perhaps a GNOSTIC UNCOUPLING would be more appropriate for this account. Nevertheless, it would also still predict the impossibility of hedged bullshitting. As I will soon outline in the following §1.3, hedging is essentially built off considerations relating to the KNOWLEDGE NORM OF ASSERTION, thus the inconsistency between Gjelsvik's theory of bullshit, hedging, and the HB THESIS would only be exacerbated.

11 Although it is worth noting that these Frankfurt+ accounts tend to only be relevant in explicit discussions of bullshitting simpliciter—i.e., bullshitting conceptual analysis. When bullshitting is merely applied to other areas, Frankfurt's account is almost invariably the one employed. For recent examples of this phenomenon, see Gibbons (2023) where bullshit is linked with politics and Hicks, Humphries, and Slater (2024) where bullshitting is connected to large language models like ChatGPT. I note this because what is paramount is that Frankfurt's bullshit is clearly beholden to ALETHIC UNCOUPLING.

12 See Carson (2010, 62) for his case in this vein of a student taking an exam who does not care about what is true *generally* but does care about doing well in the exam so makes sure to only write true things.

and thus is also committed to ALETHIC UNCOUPLING.

In sum, the key theory of bullshit of Frankfurt's, and what I take to be its successor from Stokke and Fallis, are committed to this idea that bullshit assertions are uncoupled from the alethic. Now we have two thirds of our pieces in place—viz., ALETHIC UNCOUPLING and the HB THESIS. Next, we turn to the ALETHIC CONNECTION component of hedging.

1.3 HEDGING

We earlier saw some classic examples of hedges[13] from the hedged bullshitting cases (1)–(5) above such as the verb in *I think* from (3),[14] the adverb of *probably* in (4), and the auxiliary of *might* in (5). (1) and (2) are somewhat less standard—the adjective of *certain* and explicit *I don't know* flag respectively—but still hedge their content. There are countless other epistemic terms that can similarly be used to hedge by showing the attitude or source of evidence of the hedger but I will not go through any more here.[15]

The standard line when it comes to hedging is that one hedges by including such epistemic terms in their assertions in some way which then attenuates the strength with which one put forward the content of the declarative. This is called the *strength intuition* by van Elswyk and Willard-Kyle (forthcoming, 5) or the *weakness effect* in van Elswyk (2024). Take the intuitive difference in presentational strength between (6b) and (6c):

6. (a) Who ate the ragù?
   (b) Kenji ate the ragù.
   (c) Kenji ate the ragù, I think.

(6b) and (6c) both provide the same answer to (6a), but we consider the former more forceful than the latter in some way or another. The natural way to parse this is that, by adding the epistemic qualifier of 'I think', (6c) explicitly flags a latent possibility of incorrectness that is not present in (6b).

13 Following the orthodoxy in philosophy (Benton 2011; Benton and van Elswyk 2020; Betz-Richman 2022; van Elswyk and Sapir 2021; van Elswyk and Willard-Kyle, forthcoming; van Elswyk 2018, 2023, 2024), I focus solely on a specific type of hedging, sometimes known as *speech act hedging* (Fraser 2010) or *shield hedging* (McCready 2014). For the other main type, *propositional hedging*, used for fudging categorisations, such as in 'It was *technically* an own goal', see Lakoff (1973).

14 Although it is in the matrix position here, *I think* in the parenthetical—for instance, in "*p*, I think"—is *the* paradigmatic hedged assertion. (3) is also in fact most probably a *dual hedge* (Salager-Meyer 1997) due to the further detail of "…*I may* have figured out why"—this will come up again in §1.6.1. See also McCready (2014, 41) for more on compound hedges.

15 None of the hedge terms that will feature in this chapter are controversial in any way so giving precise and exhaustive lists of what is or is not a hedge term is unnecessary. Nevertheless, see Fraser (2010, 23–25) for an impressively large list of various hedges and their categories. van Elswyk (2024, sec. 3) also discusses a vast selection of epistemic term hedges.

It is also important to note that the mere presence of a term that can be used to hedge does not guarantee a hedge interpretation—this *weakness effect/strength intuition* is not exhausted by the compositional semantics. Compare the following with (6) above:

7. (a) Who probably ate the ragù?
   (b) Kenji probably ate the ragù.

Had (7b) been in answer to (6a), we would have another hedged interpretation of *Kenji ate the ragù*. However, because (7a) is specifically asking about *what is probable*, the use of the adverb *probably* in (7b) no longer hedges—it answers the question directly without qualification. So, the same term can function as a hedge or not depending on details surrounding the context of the discourse—generally whether the question already contains the epistemic term—not merely the compositional semantics.[16] And when a term *does* hedge, it weakens the presentational strength of the proffered content.

A natural question at this point might be: what exactly does a hedge weaken from? The answer is *knowledge*, as van Elswyk writes: "Knowledge sets the threshold for hedging." (2023, 344) Bare or unqualified assertions represent knowledge (van Elswyk 2021) as put forward by the KNOWLEDGE NORM OF ASSERTION—viz., one must: assert p only if one knows that p.[17] What this means is that a term hedges when it signals that a speaker *lacks knowledge*—the difference between (6b) and (6c) is that the former signals that the speaker *knows* that Kenji ate the ragù, whereas the latter signals that they occupy a position lower than knowledge on whether Kenji ate the ragù, namely that they merely *think* so (but they could be wrong). Hedging's ALETHIC CONNECTION should be starting to become apparent—indeed, it is not obvious how hedging would even work as a linguistic device if it was not explicitly connecting to possible truth/falsity.

A variety of linguistic data is adduced to support hedging suspending the knowledge signal[18]—I shall briefly note one as it will be relevant for our purposes later. *Challenge data* further establishes the idea that bare assertions signal knowledge while hedging signals something lower. Consider the following two discourses:

8. (a) Kenji ate the ragù.
   (b) How do you know that?

9. (a) Kenji ate the ragù, I think.
   (b) # How do you know that?

16 For more on this see van Elswyk (2024, sec. 3).
17 See, of course, Williamson (1996, 2000). For further discussion among many, many more, see DeRose (2002), Benton (2012), and Simion (2016).
18 See van Elswyk (2024, 5) for discussion of hedging related to Moorean conjunctions.

(c) Why do you think that?

The factive challenge in (8b) is appropriate in response to the bare declarative of (8a) because the speaker was signalling that they *know* that Kenji ate the ragù. In (9b), the same challenge is defective because the speaker in (9a) was not signalling knowledge (they only said they *thought so*), hence the challenge in (9c) is more natural.

We are now in a position to plainly lay out the essential characteristic of hedging:

> KNOWLEDGE CHARACTERISTIC
> Hedging signals that the speaker occupies a position lower than knowledge towards the asserted declarative.

Notice that ALETHIC CONNECTION is contained under this key aspect insofar as such signalling of below knowledge is only possible through an explicit connecting of the assertion to the alethic—viz., to truth values and thus possible falsity or correctness. In fact, ALETHIC CONNECTION is almost just a recasting of van Elswyk's account of the etiological function of hedging: to alert hearers to positional risk (van Elswyk 2023, sec. 3). To briefly explain, hedging signals to hearers that the speaker does not know the assertion in play, and lets them know precisely what position the speaker occupies in relation to the content (be that *thinking it*, *believing it*, *it being possible for them*, etc.) so that the hearer can have an appropriate doxastic response to the testimony. This signalling only works because the hedge explicitly connects to and flags the alethic (in other words, explicitly raises the possibility of falsity).

Importantly, ALETHIC CONNECTION stands irrespective of the speaker's *goals* when they hedge—the work is done entirely by the hedge terms themselves and the surrounding discourse context. For instance, one might think that a deceptive hedge—e.g., my saying "*p*, I think" when I actually believe that *p* is false—is not connected to the alethic. This would be incorrect, however. Even if the hedge is *inaccurate* or *deceptive*, so long as there is a hedging interpretation available (in other words, *not* like discourse (7) above) then it *explicitly* flags its potential falsity (its *alethic connection*) simply by virtue of its presence. As Benton and van Elswyk write: "speakers who opt for hedging their assertions represent themselves as less confident given that they could have, but did not, unqualifiedly assert," (2020, 252) or as van Elswyk and Sapir remind us: "Speakers who hedge choose not to represent knowledge." (2021, 5851) The speaker's intentions, actual epistemic position, and so on, are immaterial to what the hedge *represents* to the hearer about the speaker's epistemic position—this straightforwardly aligns with the position represented by the hedge.[19] Hence, the characteristic of ALETHIC CONNECTION. And, of course, this is all entirely compatible with the standard case that, "hedging is usually accompanied by a belief that the content of the hedged statement is not known." (5851)

19 See van Elswyk (2023, 354) for his "P-norm requirements" table. To give an example: *I guess that* p signals that p *is being guessed*—this is true even if the speaker knows that *p* or thinks it is false, etc.

This KNOWLEDGE CHARACTERISTIC/ALETHIC CONNECTION often results in normative effects.[20] Let us close out this exegetical section by considering two such effects:[21]

> RESPONSIBILITY EFFECT
> Hedging makes a speaker less responsible for the content of their assertion.
>
> REPUTATION MANAGEMENT
> Hedging can exempt the hedged testimony from affecting the speaker's reputation.

We have in fact already seen an instance of the RESPONSIBILITY EFFECT in action: the *challenge data* above. Through hedging, the speaker in (9a) is not responsible for answering as strong a challenge as (9b) is, hence the impropriety of such a challenge, and only has to answer the weaker challenge asking *why they think so*. The thought is that answering why one *knows* something is more arduous than answering why one merely *thinks* something—presumably because explaining purported knowledge of the case generally requires greater justification than explaining why one just thinks something is the case.[22] This reduction in responsibility can also be made salient in a different way; suppose Kenji actually did not eat the ragù; Shabu the dog was the real culprit. Intuitively, the speaker in (8a) who spoke sans qualification that Kenji ate the ragù (in other words, signalled that they *knew* Kenji ate the ragù) is up for more opprobrium than the speaker in (9a) who merely signalled that they *think* Kenji may be the perpetrator—this is because through hedging the speaker is less committed to (or *less responsible for*) the content of the assertion (being true).

REPUTATION MANAGEMENT can at times be hard to entirely demarcate from the former effect but there is a sense in which it has broader scope. The idea is that the speaker is not just shielded from some responsibility if wrong (as was just covered), they are also shielded from (some) approbation if *right*. Consider the following: suppose now that Kenji did in fact eat the ragù. One gets the impression that the speaker in (8a) *did better* by signalling knowledge correctly than the speaker who hedged (but

20 I say "often" here and not "always" because although a hedge interpretation always suspends the knowledge signal, it does not always reduce one's responsibility (viz., the first normative effect to come). To give a quick example: a doctor telling a patient, "I *think* this is the correct medicine," *does* signal that they do not *know* whether it is the right drug, but it obviously would not reduce their responsibility (qua doctor) were they ultimately wrong.

21 The two effects to be outlined are where the majority of hedging discussion (outside philosophy) has been focused. See Brown and Levinson (1987) and Fraser (2010) for the first to come, and McCready (2014) for the second. Both are also discussed in van Elswyk (2023, 345–350) to the end of rejecting them as the potential etiological function of hedging.

22 One could felicitously issue the weaker challenge of *Why do you think that?* in response to the unqualified assertion in (8a) but the key point here is that the stronger challenges of *How do you know that?* or *You don't know that.* will always be inappropriate in response to a hedged declarative.

was still correct) in (9a). The connection between the hedging relevant in this chapter and the hedging referred to in the idiom *hedging your bets* is made salient here; hedging reduces risk but also reduces reward. Just as letting a bet ride without hedging it would result in a greater reward if it wins, but with greater risk as there is no protection if it loses, so too does asserting without qualification result in greater reputational reward if the assertion is true but with greater reputational risk as there is no *below-knowledge* signal shield if it is false.

In sum, hedging an assertion signals that one does not know it—in other words, it explicitly connects it to the alethic and flags possible falsity. In turn, this can have responsibility-insulating effects. These will be relevant later when we look at the consequences of consummate bullshitting.

1.4 THE BONA FIDES OBJECTIONS

Let us briefly pause and take stock. We opened with some examples of sentences that were prima facie both bullshit and hedged—cases (1)–(5) of hedged bullshitting. From their existence, we drew out the HB THESIS which made the weak ability claim that one can hedge their bullshit (i.e., bullshit and hedge at the same time). I then spent the last two sections showing that, first, standard theories of bullshit (Frankfurt 2005; Stokke and Fallis 2017) are committed to the idea that bullshit assertions are uncoupled from the alethic (ALETHIC UNCOUPLING) and, second, that hedged assertions are essentially and explicitly connected *to* the alethic (ALETHIC CONNECTION). Pulling all this together, we have an inconsistency—standard theories of bullshit predict that hedged bullshitting is impossible; the HB THESIS and (1)–(5) say otherwise. So, something has to give, be that ALETHIC UNCOUPLING, ALETHIC CONNECTION, or the HB THESIS. I will propose that we jettison the standard accounts of bullshit and thus ALETHIC UNCOUPLING. First, however, I will consider some responses from my opponent: one or both of hedging or bullshitting are not *really present* in the cases, they say, so we can reject the HB THESIS. I call these the *bona fides objections*.

The line is as follows: granted one cannot hedge their bullshit for the alethic reasons outlined. Nevertheless, there is not an inconsistency here because (1)–(5) are not actually hedged or not really bullshit, thus the HB THESIS was derived from error. We can lay out the potential avenues of these objections as follows:

In actuality:	**Bullshitting**	**Non-bullshitting**
Hedging	*Hedged bullshitting*	*Hedged non-bullshitting*
Non-hedging	*Non-hedged bullshitting*	*Non-hedged non-bullshitting*

Table 1.1: *The dialectical space of the bona fides objections*

I will now show that none of *hedged non-bullshitting*, *non-hedged bullshitting*, nor *non-hedged non-bullshitting* make for plausible explanations of the cases, thus (1)–(5) remain as bona fide examples of hedged bullshitting and the HB THESIS stands.

1.4.1 Non-hedged bullshitting

The first of the three alternative explanations we will consider is the *non-hedged bullshitting* response. The idea is that the cases are genuinely instances of bullshitting but not proper instances of hedging—it is merely a facsimile of a hedge or perhaps the hedge is just a continuation of the bullshit. It is not a *bona fide* hedge. Therefore, the HB THESIS can be rejected as one is not actually hedging their bullshit assertion, they are merely bullshitting simpliciter.

This does not make for an adequate explanation of our cases, however. Consider again (5) and an alternative, *unqualified* variation, (5′):

5. Trump might break 100,000 attendees tonight. Absolutely insane.

5′. Trump will break 100,000 attendees tonight. Absolutely insane.

(5) is, according to this response, *not* hedged bullshitting, but rather, say, *fake-hedged* bullshitting. (5′) is simply a (bare) bullshit assertion. Therefore, per this objection, both of these sentences are classified as instances of *unqualified bullshitting* about the same content (# of attendees at the rally). Evidently, however, they are *not* the same, and the same content does not look like it is being presented with the same force.

Nevertheless, the objection might continue, as already noted, the hedge is bogus, fake, insincere—it is *mimicking* less force or commitment—so while there is *superficially* or *syntactically* a difference between (5) and (5′), the former is not properly hedged in some way or another. Perhaps the speaker in question does not merely *think it is possible* that numbers break 100,000—that is, for example, they think it is guaranteed, or they do not believe it at all. So, the hedge is deceptive (in some way), and this in turn means our apparent hedged bullshitting cases are in actual fact cases of *non-hedged* bullshitting.

I have no doubt that the hedges in many cases like this can be deceptive or misleading. Nevertheless, that does not make this response a correct one. It is mistaken

simply because of what a hedge represents and that "speakers who choose to hedge choose not to represent knowledge." (van Elswyk and Sapir 2021, 5851) Recall the discussion in §1.3, the speaker's goals, intentions, or actual epistemic position towards the content do not affect the fact that a hedge (interpretation) explicitly connects the assertion to the alethic—irrespective of the hedge's possible mendacity. The fact that the hedge may be deceptive or bullshit in some way does nothing to adulterate what the speaker intended to represent to the hearer.[23]

What this response really requires to make the case that (1)–(5) are actually *non-hedged* is that there is no hedging interpretation present despite the use of hedge terms, thus ensuring an *unqualified* characterisation (a la discourse (7)). Nevertheless, linguistic data can be fielded to show that this not the case either.

Take (1)–(4) and challenge them with our factive challenges of either *You don't know that.* or *How do you know that?* and note the results: they clearly sound defective. (3) in particular is unavoidably egregious considering one follows *I don't know the answer to that.* with the "challenge" *You don't know that.* If (1)–(4) were truly *non-hedged* then these challenges should be felicitous, which they evidently are not, thus we have good evidence that we are not improperly characterising them as hedged. In fact, (5) is the only case that I do not immediately get a clear read on, but we can even explain it.

Consider the following discourse:

10. (a) Might Trump break 100,000 attendees tonight?
    (b) Trump might break 100,000 attendees tonight.
    (c) ? You don't know that.

There is no hedging interpretation here in (10b) (due to the question and answer both being about *what might be the case*), thus the challenge in (10c) is, to my mind, natural because the speaker in (10b) is representing knowledge about *what might be the case*. Nevertheless, I am happy to allow that (10c)'s felicity is at least somewhat up in the air as all I need is that this response is not as noticeably degraded as the following discourse's (11c):

11. (a) How many attendees will Trump break tonight?
    (b) Trump might break 100,000 attendees tonight.
    (c) # You don't know that.

(11b) is clearly hedged so (11c) is obviously defective. I argue that we have good reason to think that the *implicit* QUD in (5) was not (10a) but rather (11a), meaning (5) is correctly diagnosed as being hedged. This is so for two reasons. First, (11a) is a far

23 Assuming the hearer does not know it is deceptive or bullshit, that is. I will set aside this thought as there is not much of interest here if everyone already knows when they are being bullshitted.

more natural implicit QUD—there is nothing in (5) that implies the speaker would be talking about epistemic possibilities. Second, the added judgement of "Absolutely insane" presented in (5). If, somehow, the speaker took themselves to be only talking about epistemic possibilities (viz., what *might* be the case), then this opinion would be rather out of place—presumably affirmative answers about weak epistemic possibilities are rarely "absolutely insane". Assuming the implicit QUD of (11a), on the other hand, such a statement makes sense: an absolutely insane amount of attendees is 100,000 because that is ludicrously large (as evidenced by the fact that the actual total was not even close). Therefore, (11) is most plausibly the (implicit) discourse in play in (5), which has an inappropriate knowledge challenge, thus providing good evidence that (5) is hedged. So, all five of our hedged bullshitting cases pass the tests used to check for hedging, and thus we can reject this objection's *non-hedged bullshitting* characterisation.

1.4.2 Hedged non-bullshitting

At this point, my opponent might pivot and cry Gift of the Magi. Perhaps I have oversold the hedging and so lost the bullshitting. That is, the examples are indeed bona fide *hedged assertions*, it is the *bullshitting* that is absent; perhaps the hedge somehow washes away the bullshit.

I do not think this response holds water either. Aside from the base fact that I think there is a strong intuitive pull that the cases *just are* instances of bullshitting, the main issue stemming from this line of objection is that, if not bullshit, the cases must be *something else*, and there are no obvious or particularly compelling alternatives. The available options are that the cases are "upgraded" from (prima facie) hedged bullshit to *ordinary hedged declaratives*, or "downgraded" to *hedged lies*.[24]

Starting with the former, it quickly runs into the issue that, given their new status as ordinary hedged assertions, we now lack the normative resources to criticise them robustly—despite their intuitive perniciousness. If (1)–(5) are no longer instances of bullshitting then they plausibly just come out as *permissible* hedged declaratives insofar as, for example, per (1), it does seem true that we cannot be *certain* that these events are connected, nor be *certain* of the opposite. This permissibility becomes even more salient if we make the hedge term one of the weaker varieties. Consider this alternative (3):

3′. Google's Gemini AI project doesn't want to display images of white men—even historical figures. *It is possible that* I may have figured out why.

24 Some might be immediately sceptical of such transformation. For example, Betz-Richman writes: "It would be surprising if [a hedge] could magically transform the status of [an untruthful bare declarative] from a lie into some other category of speech." (2022, 5) Presumably, he would feel the same about a hedge transforming a bare bullshit assertion into some other category of speech. Nonetheless, I will give this *hedged non-bullshitting* line a full treatment.

Something's being possible for someone is a low epistemic bar to clear, so it would be difficult to argue this is false (i.e., a lie) or misleading even. When we are still considering it as *bullshit* the fact that it might ultimately be true (in this sort of minimal way) does not affect the fact that we can dismiss it as bullshit or criticise the speaker as a bullshitter. However, if we lose these normative grounds of its being bullshitting as this version of the *hedged non-bullshitting* response requires, then there are no easy grounds to inject opprobrium. This is a poor consequence of the response.

Nevertheless, there still remains the *downgrade* option. The perniciousness spectrum of deception has it going: lying, bullshitting, then misleading, with lying being the worst. As Webber puts it: "Honesty is the best policy. But if you must depart from it, then you should mislead first, bullshit second, and lie only as a last resort." (2013, 659) So, this version of the response goes, the addition of a hedge to bullshit somehow shunts it down a category to a *hedged lie*.[25]

This route is not promising either. Recall the account of lying mentioned in §1.2:

> LYING DEFINITION
> To make a believed-false statement to another person with the intention that the other person believe that statement to be true. (Mahon 2016)

In most of our cases, it does not look like any of the speakers would believe that what they are saying is false—in fact, the presence of the hedge term in some sense signals a care to nominally track truth. As noted just above, if the hedge is an epistemically weaker one (like *possibly*), then there would be no reason to expect the speaker to believe what they are saying is false—indeed it would be very strange if they did—because what they are saying is very likely to be true. To reuse (8′), it would be rather odd for the speaker to in fact believe that it is *impossible* for them to have figured out what is going on. They may well believe that they are possibly incorrect, but that is of course entirely congruent with a paired belief that it is also possible that they *are* correct. Importantly for our response to this *hedged non-bullshitting* line, none of this is commensurate with the speaker in question *lying*.

Of course, the LYING DEFINITION above is not the sole account of lying. Nevertheless, the other main player in the literature, the KNOWLEDGE DEFINITION OF LYING (Benton 2018) involves the speaker *knowing* that what they say is false, which is presumably impossible in many of our cases insofar as they are either future-directed or epistemically isolated to such a degree that *knowledge* about their truth value is far too difficult to achieve. Moreover, this is assuming all the statements turn out false—which is not guaranteed.

A final point of conjecture: (1)–(5) come from speakers who I take to be inveterate bullshitters. If we assume in these cases that they are in fact *lying*, then that would presumably entail that they are actually inveterate (pathological?) *liars*. My guess is

25 The existence of which is convincingly defended by Betz-Richman (2022).

that it is psychologically far easier to merely bullshit constantly than lie constantly simply because continually asserting falsehoods knowingly is far more difficult than just speaking with abandon without worrying about anything as juvenile or boring as the truth. Moreover, the speakers seem to consider themselves intelligent truth seekers, so the cases being bullshit over lies I just take to be more plausible. There is some empirical data in support of this, such as Petrocelli (2018) who writes: "Thus, people may generally underestimate the degree to which they engage in bullshitting," (256) whereas one cannot lie unknowingly. In sum, the *hedged non-bullshitting* response fails. It does not capture the cases well while also being an intuitively unpalatable explanation of events.

1.4.3 Non-hedged non-bullshitting

Our final option is that the cases are in fact neither hedging *nor* bullshitting, but something else entirely. This option has the greatest dialectical burden to shoulder insofar as it has to first overcome the arguments fielded above for why the cases are hedged *and* why they are bullshit, *and then* has to tell a good story explaining how both elements disappear to form an entirely different category of speech. All this to say, I do not think this is a particularly plausible path to take, so we will not consider this option.

1.4.4 Hedged bullshitting

So, all the alternatives to hedged bullshitting are either implausible or unpalatable. (1)–(5) remain *hedged bullshitting* and thus we maintain the HB THESIS. Therefore, we still have the inconsistency between our three main claims insofar as I have argued the HB THESIS is true but the standard theories of bullshit predict its falsity. Therefore, it is time to flip the explanation and reject the ALETHIC ACCOUNTS and their ALETHIC UNCOUPLING.[26]

1.5 A DIFFERENT KIND OF BULLSHIT

Standard theories of bullshit, what I termed ALETHIC ACCOUNTS, are committed to ALETHIC UNCOUPLING—that is, that bullshit assertions are uncoupled from the alethic. Combined with the HB THESIS and hedging's ALETHIC CONNECTION, we got an inconsistency. If, however, we could give an alternative, compelling, *non-alethic* account of bullshit, this would dissolve the tension as there would be no ALETHIC

26 I will not consider the move that we reject ALETHIC CONNECTION for two reasons: first, I think ALETHIC CONNECTION and the KNOWLEDGE CHARACTERISTIC it is drawn from are on nigh unimpeachable ground; second, I already have a good explanation for events in the following section if we jettison ALETHIC UNCOUPLING.

UNCOUPLING to contravene the combination of ALETHIC CONNECTION and the HB THESIS. Well, it just so happens that there is such an account. Enter Cova (2024).

Cova distinguishes between *process-based accounts* and *output-based accounts* of bullshitting.[27] The ALETHIC ACCOUNTS we considered earlier were all *process-based* as they make bullshitting about the properties (or lack thereof) of the assertion and thus some piece of bullshit is just whatever is created by this process. Cova's new account is primarily *output-based* which means it focuses on the statements themselves and builds the analysis on their features. The essence of Cova's account, pace ALETHIC UNCOUPLING, is that bullshit is something that looks interesting at first sight but is always deflatable when placed under close(r) scrutiny. More precisely:

> What makes a given claim C bullshit is that (i) though C is presented as or appears at first sight as making an interesting contribution to a certain inquiry, (ii) C would turn out, on closer inspection by a minimally competent inquirer,[28] to make a much less interesting contribution. (587)

Cova does a lot of work to show that this new account is more extensionally adequate than any extant ones mainly through examples of what he calls "truth-tracking bullshit" (577–578), among other virtues.[29] None of these literature-specific details are important for this chapter but it is a boon that this account nicely explains all the examples of hedged bullshitting employed throughout. For instance, the claim from (5) that Trump *might* break 100,000 attendees looks very exciting and interesting at first glance, then comes the important clarification that the picture attached is of 10,000 people and the eventual total will be 21,000. The claim in (2) that the truth is possibly being withheld from us posits exciting and interesting conspiracies and subterfuge, except of course in reality this is not happening, and so on for the other cases.

This is an account of *bullshit*, however, not an account of *bullshitting*, and it does seem that the latter is what I am more interested in here for our discussion of hedged bullshitting. Importantly, Cova does also give an account of bullshitting more in keeping with the process-based focus of the past. In essence, Cova takes bullshitting to be the activity or performance of giving off an undeserved affect to someone. Specifically:

27 This distinction is credited to G. A. Cohen (2002).

28 One might worry that "minimally competent inquirer" is too weak as some bullshit can presumably be complicated or well disguised such that it would take, say, a *fairly competent* inquirer to note its bullshittingness. Cova anticipates such a worry and clarifies that the "minimally competent inquirer" is domain-relative, using an example from physics (587-588) to explain, the idea being that some complicated bullshit in some domain only looks so to the laity, not to a minimally competent inquirer *within that domain*.

29 A central example employed by Cova is that of a menu in a fancy restaurant that uses purple prose to pretentiously describe fish and potatoes. Cova takes this to be a paradigmatic case of bullshit that cannot be countenanced by any extant accounts because it carefully tracks the truth in all respects and so is antithetical to ALETHIC UNCOUPLING (although he does not use this terminology). He also notes that his account better aligns with seminal work outside philosophy such as anthropologist David Graeber's *Bullshit Jobs* (2018).

> X is bullshitting when X engages in a communicative act C but is more concerned with the general (affective) impression their act will have on a given target T, than about the particular propositions they will get T to endorse AND X is aware that X has no good reason to think that the impression X is trying to convey is warranted by reasons presented in or hinted at by his communicative act C. (Cova 2024, 594)

This is rather wordy, and the finer details are not paramount here but what is important is this idea of giving a certain impression to someone that is ultimately not appropriate given what they are communicating. Again, we can see that this well explains our cases (and dovetails nicely with the account of *bullshit*). The speaker in (1) wants to seem like a genius of geopolitics reading between the lines and spotting deep patterns, the person putting forward (3) wants to give the impression of a hardcore investigation into the ins and outs LLM coding, *mutatis mutandis* for the rest. Admittedly, it is perhaps not very surprising that Cova's account gets all the cases right; after all, it purports to be the most extensionally adequate account on the market, and I have been insistent throughout that the cases employed in this chapter are indeed bona fide cases of bullshitting.

Let us take stock again. I have now explained an alternative theory of bullshit from the ALETHIC ACCOUNTS which characterises all our central cases as ones of bullshitting. Cova's account makes no claims of alethic disconnection and so we can jettison ALETHIC ACCOUNTS and their ALETHIC UNCOUPLING commitment that rejected the possibility of hedged bullshitting. Thus, we can maintain the HB THESIS and hedging's ALETHIC CONNECTION without any inconsistency as Cova's account has no quarrel with ALETHIC CONNECTION. And so we get (1)–(5) as bona fide instances of the novel phenomenon of hedged bullshitting without any incoherence.

The first key thesis of this chapter is now complete: I have demonstrated that hedged bullshitting is a genuine phenomenon and thus we ought to endorse Cova's alternative account of bullshit to countenance it. However, we are not done yet. Indeed, one might feel there remain some gaps I ought to fill: first, I said Cova's account can *allow* for hedged bullshitting, but we might want more than that insofar as it would certainly be a boon if his theory could also *explain* it; second, we might wonder *why* one would ever hedge their bullshit. What actually happens when one hedges their bullshit and what is the point in it? In the following and final section, I will answer these questions.

1.6 HEDGED BULLSHITTING

1.6.1 Understanding hedged bullshitting

Let us start with the first gap fielded above, the explaining of hedged bullshitting. After all, while Cova's account might not immediately disqualify hedged bullshitting from being possible the way the ALETHIC ACCOUNTS did, if it cannot do much to explain the phenomenon, it would be a mark against the theory's card. Nevertheless, I will now argue that we get a full and satisfactory explanation of hedged bullshitting under this non-alethic account.

Recall that Cova places the core of *bullshitting* in this conveying of an undeserved affective impression on one's target. With respect to this aim, the hedges featuring in (1)–(5) only enhance the performance of this bullshitting. Our bullshitters in the cases all essentially want to come across as intelligent and legitimate commentators on their chosen subject—what better way to fake that than direct references to possible falsity, admissions of intellectual humility,[30] a careful and specific dearth of *any* knowledge claims displaying a clear awareness of their own and indeed *human* fallibility. These speakers want to seem like conscientious inquirers who respect the truth and are always willing to admit that they might be wrong; this is all part of the deception. They think deeply and meticulously and even then put forward their ideas with provisos that explicitly note that they *cannot be certain*, that they only *think they may* know what is going on, that they in fact *do not know all the answers*.[31] Empirical data supports this. Multiple studies have found that hearers trust speakers more when they admit of fallibility in what they say when they are indeed wrong (Sah, Moore, and MacCoun 2013; Tenney et al. 2007; Tenney, Spellman, and MacCoun 2008). And bullshitters are presumably going to be wrong more often than most considering their priorities, so hedging is a vital tool in their employ.[32]

There are also some what we could call broader theoretical *coherence benefits* in this newfound harmony between hedging and bullshitting. Consider the following discourse, imagining the questioner to be a journalist, the answerer a politician, and X some social policy:

30 See Hazlett (2020) for a discussion of *false* intellectual humility. I take this to be a related area to my discussion in this chapter.

31 Cf. Cassam (2019a). Although he is discussing conspiracy theorists, who are not necessarily bullshitting (and nor are bullshitters necessarily conspiracy theorists), there are some interesting similarities between the cases in this chapter and some of what he discusses. For instance, "Conspiracy Theorists who are quick to denounce mainstream academia for rejecting their theories nevertheless crave academic respectability. They set up pseudo-academic journals for the study of this or that alleged conspiracy and trumpet their PhDs, whatever their subject. They have a particular fondness for footnotes… because it creates the impression that his theories are the product of reliable research into trustworthy sources."(20) See Chapter 5 for an in-depth discussion of the philosophy of conspiracy theories.

32 The utility of hedging when bullshitting should be becoming clear. The following section will fully drive this home.

12. (a) Will the budget stretch to include X?

    (b) It would seem somewhat unlikely that X won't factor in.

(12b) looks like a classic instance of a politician bullshitting insofar as they are saying a lot of words that do not really say much at all, and being very careful to not assert anything too strongly—hence the quadruple hedge (Salager-Meyer 1997) of *It would seem somewhat unlikely*. So here we have another example where hedging and bullshitting are not in tension as the ALETHIC ACCOUNTS would predict but actually closely tied together—which is, again, easily explained on Cova's account; (12b) purports to genuinely answer the question and the hedges contribute to perhaps make it appear more interesting or satisfying but it ultimately quails when put under any kind of scrutiny. This is paradigmatic question-evasion bullshit from the politician and hedges are often explicitly described as evasion strategies (Fraser 2010, 27–28).

A final interesting titbit relates to *advertising*. Frankfurt's (2005) original work takes a lot of advertising to be paradigmatic instantiations of bullshitting; Johnson (2010) explicitly writes about bullshit and advertising; Cova discusses advertising in relation to bullshitting at various points (2024, 593)—in essence, if your account of bullshit did not or could not include any advertisements then it would undoubtedly be insufficient. Now, in her *Reliability in Pragmatics*, McCready (2014) throughout links hedging, disclaimers, and *advertising*, for example:

> The speaker's goal in hedging is to avoid being held responsible for the content of his utterance if it proves to be false, just as the goal of a disclaimer in an advertisement is to avoid being held responsible if the actual product is less satisfactory than the advertisement makes it out to be. (39)

She seems to take them to be essentially branches of the same tree and the comparisons and analogues are compelling. Of course, these coherence claims are ultimately somewhat vague but the connections between hedging and bullshitting here are genuinely interesting and more grist for my mill insofar as I think it lends credence to the thought that these are important and plausibly paradigmatic cases for the philosophy of bullshit. At the very least there are pro tanto reasons to think that hedging and bullshitting are more closely related than has previously been appreciated and, again, *only* Cova's non-alethic account is at all placed to explain such a connection—and explain it does.

1.6.2 Getting away with bullshitting

Still, we might wonder if the juice is really worth the squeeze when it comes to hedging one's bullshit. After all, bullshitters are often in the game of deceiving people and pushing certain political (or otherwise) agendas in addition to the financial incentives I noted right at the beginning of §1.1. So, they presumably want people to

believe the bullshit they say, not merely *see* it. As van Elswyk notes, however: "Belief is warranted by all unhedged testimony, but it is only warranted by some instances of hedged testimony." (2023, 355) Indeed, out of our five hedged bullshitting examples, plausibly only (3) warrants full belief from the hearer, the rest may just permit some changes in credences.[33] Plausibly, then, one might think the bullshitters are actually doing their targets a favour insofar as many hearers will often not take up the bullshit beliefs when the assertions are hedged.

Nevertheless, I want to wholly resist this proposal. In fact, in this final part of the chapter, I will argue for the exact *opposite*—that hedged bullshitting is often far more dangerous than its unqualified counterpart. A large part of this relates to the way the normative effects of hedging are uniquely positioned to assuage key vulnerabilities associated with consummate bullshitting—the details of which are to come in the following section. Beforehand, however, let me respond to the objection raised directly above about hedged bullshit not resulting in people believing the bullshit—and thus basically being a blunder on behalf of the bullshitter with an agenda.

This sort of response is misguided. The key reason why is that deceptive agendas are rarely about getting targets to *directly* inculcate falsehoods.[34] They are more insidious than that and often an exercise in seeding doubt and defeating one's (true) extant beliefs (at least initially).[35] As Mikkel Gerken writes about the SALIENT ALTERNATIVE EFFECT when it comes to people's (loss of) knowledge of climate change: "Roughly, this is people's disinclination to accept ascriptions of knowledge in the face of contextually salient error possibilities." (2022, 149–150) The thought here is that irrespective of our hedged bullshitters' deliberate avoidance of any knowledge claims, they can still make salient their bullshit error possibilities such as the World Health Organisation being a nefarious institution, or there being a conspiracy where the truth is hidden from the laity, and so on. And hedged bullshit certainly carries the doxastic weight to cause this effect: "hedged testimony… can make the hearer newly aware of what epistemic possibilities are warranted for them." (van Elswyk 2023, 358) So, this objection that hedging one's bullshit incapacitates its potential for deception is mistaken. Moreover, as I will now argue, the inveterate bullshitter usually has far bigger problems than merely getting their deceptive agendas across, and the ameliorative work hedging does here for the bullshitter is striking and pays dividends that far exceed any potential doxastic warrant trade-offs.

33 For more on the warrant component of hedged testimony and how different hedge terms warrant different credences/beliefs, see van Elswyk (2023, sec. 4.3).

34 See §3.2.3 in the third chapter for a more detailed discussion of the following point.

35 For the most infamous cases of this exact strategy being employed in relation to smoking tobacco and global warning, see Oreskes and Conway (2011).

1.6.2.1 A tension

At the heart of the bullshitting literature in philosophy, there lies an unacknowledged tension. I will cash it out here in three distinct elements:

(i) Bullshitting is very dangerous to society.

As Frankfurt said: "bullshit is a greater enemy to truth than lies are." (2005, 61) Cova took it as a serious desideratum of accounts of bullshit that they countenance its dangers. A recent study even concurred, finding that, "bullshit appears to have a more potent impact on beliefs about what is true, and one's own attitudes, when the very same information comes from a bullshitter than a liar." (Petrocelli, Silverman, and Shang 2023, 9616) On the other hand, however:

(ii) Sporadic bullshitting is likely insignificant. Rather, it is constant bullshitting that foments this danger.

Frankfurt frequently notes this, such as when he writes:

> Through *excessive indulgence* in [bullshitting], which involves making assertions without paying attention to anything except what it suits one to say, a person's normal habit of attending to the ways things are may become attenuated or lost. (Frankfurt 2005, 60) [my emphasis]

Or:

> In contrast, indifference to the truth is extremely dangerous. The conduct of civilized life, and the vitality of the institutions that are indispensable to it, depend very fundamentally on respect for the distinction between the true and the false. (Frankfurt 2002, 343)[36]

Frankfurt does not mean that that one time you bullshitted your way through an inconsequential conversation resulted in the beginning of the end for the conduct of civilised life. Rather, the point is that if a broad swathe of society were to go about *all* or *most* of their interactions in this bullshit way and fully inculcate a standard process of bullshitting across the board, then the societal harms would be apparent. Lastly, however:

(iii) One tends to only get away with sporadic instances of bullshitting. Consummate bullshitting almost inevitably results in being challenged and having one's reputation tarnished (as a bullshitter).

36 The quote here involves disconnection from truth because Frankfurt's bullshitting involves ALETHIC UNCOUPLING. Of course, we have rejected this way of understanding bullshit but that does not mean I cannot employ his wider claims about the phenomenon's societal effects. It is also interesting to note that the quote preceding this one is commensurate with the Cova account I have endorsed here.

I take this to be an intuitive thought. If someone is repeatedly bullshitting with regards to provable matters (such as the cases discussed throughout), they will almost inevitably be caught out, and the more they are bullshitting the quicker such a discovery will be made. I also invite the reader to consider from their own experience anyone they knew/know as bullshitters and how their testimony was/is severely downgraded because of that characterisation. As Gibbons writes: "Politicians may suffer reputational harms if they are known bullshitters," and, "the extent to which one suffers reputational harms because of one's bullshitting may vary with one's ability to bullshit in a convincing manner." (Gibbons 2023, 6) Outside philosophy, there is also evidence for (iii), such as one study which proposed that, "when people are held accountable or when they expect to justify their positions to people who disagree with their attitudes—people appear to refrain from bullshitting," (Petrocelli 2018, 255) and another that found that, "bullshitting was relatively frequent when the social context lacked a cue to accountability." (Petrocelli, Watson, and Hirt 2020, 244)

We are now in a position to clearly see the tension: (i) bullshitting is purportedly dangerous, but (ii) plausibly only consummate bullshitting, and (iii) consummate bullshitting makes for rather ineffective deception due to challenges and accountability and so is not very dangerous. So, there is an issue with capturing the exact threat bullshit is supposed to—and indeed appears to (Petrocelli, Seta, and Seta 2023)—pose. I will now argue that the central phenomenon of this chapter, hedged bullshitting, is uniquely placed to address this third point. Specifically, I will argue that hedging deals with the dual problems of *being challenged* and the *negative reputation effect*. This will tell us how bullshitting can be so dangerous *and* why our bullshitters in (1)–(5) hedged their bullshit. In essence, this final part of the chapter is a guide in ...

1.6.2.2 How to hedge bullshit and influence people

Compare the following two discourses (13) and (14):

13. (a) I just wonder whether the truth is being withheld from us. I don't know the answer to that.
    (b) # How do you know that?
    (c) Why do you think that?

14. (a) The truth is being withheld from us.
    (b) How do you know that?

Note that (13a) is just our hedged bullshitting case (2). Thus, (13b) is a clear defective response as the speaker was not making a knowledge claim. Instead, to challenge (13a), the speaker has to field the weaker challenge of (13c) (or some other weaker alternative). The second discourse, (14), on the other hand, is the *unqualified* version

of the same bullshit, and thus represents knowledge, opening itself up for the factive challenge in (14b).[37]

Here is the basic idea: (13c) is a far easier challenge for the bullshitter to respond to than (14b). This phenomenon—of *How do you know that?* being somehow stronger than *Why do you think that?*—has of course been addressed before in the hedging literature:

> If a speaker anticipates a loss to their reputation by not being able to answer the challenges that would accompany unhedged testimony, hedging enables them to contribute in a manner that licenses weaker challenges and challenges that, if not answered, induce a less significant hit to their reputation. (van Elswyk 2023, 363)

So, by hedging, the bullshitter in (13a) has less expected of them in response to challenges than the bullshitter in (14a). This is presumably a significant asset considering that bullshitting likely does not easily admit of good justification or defence. However, I do not think that the challenge effects hedging has in the cases here are merely additional examples of the orthodox challenge data discussed throughout this chapter. In fact, I will now argue that the different challenges licensed have interesting normative upshots in quite a different way than they did in, for instance, discourses (8) and (9), and this can be directly attributed to the fact that the assertions being hedged here are *bullshit ones*.

The upshots I refer to can be seen when we consider and compare the potential properties of the responses the *bullshitter* can give to (13c) and (14b). It is not just that the former is "easier" to respond to than the latter, it is about the responses licensed *to such challenges*. Take discourse (14). I earlier called (14b) a "factive challenge". What I meant by that is that it immediately makes salient *knowledge* and directly asks for the story as to how one *knows* that what they are (unqualifiedly) asserting is true. In other words, it does not easily admit of *further bullshitting* because the challenged party must acknowledge what the question forces upon them—*do* they know and *how* do they know that *the truth is being withheld from us*.

In such a discourse, I propose, the speaker most likely ends up with two options to answer the challenge: a) admit they do not know the claim, retracting it altogether or at least dropping it to, say, a supposition; or b) lie. The former strikes me as a weak response that would not make for effective deception. Nevertheless, if the bullshitter wants to continue on with the asserted content that *the truth is being withheld from us* without going down route a), then they would *have to* lie because they do not know the claim at issue. So, answering *how they know it* necessarily involves saying *believed-*

37 Viewing bullshit assertions as ordinary bare assertions (in terms of normative responses permitted) is orthodoxy in the literature. For instance, Stokke and Fallis write: "bullshitting does not exempt one from the kind of commitment to what one says that characterises ordinary cases of assertion." (2017, 289)

false claims that the initial bullshitting did not require. In essence, (14) is close to being incompatible with the speaker continuing to bullshit.[38] Note that this analysis also comports with the empirical data adduced above that suggested that speakers will refrain from bullshitting when they expect to face difficult challenges and be held accountable (Petrocelli 2018; Petrocelli, Watson, and Hirt 2020).

Now compare the above with the responses licensed to (13c). *Why do you think that?* is an incredibly weak challenge compared with *How do you know that?* For a start, it is clearly not a factive challenge insofar as, while one cannot know false things, one certainly can *think* false things. Therefore, the bullshitter's response to (13c) can quite easily be more bullshit—they are not manoeuvred into a) or b) whatsoever; they do not have to admit that they do not *know* the statement in question (they were not making a knowledge claim), nor drop the assertion at all, nor lie about how they know it to be true. All they have to do to answer the challenge is say something about why they *think* the claim. Such an answer is likely commensurate with being more bullshit (and probably wholly indistinguishable from a genuinely true answer). Frankfurt had noted something similar to this before, although he did not consider the normative effects of *challenges* as I have here:

> The liar is inescapably concerned with truth-values. In order to invent a lie at all, he must think he knows what is true. And in order to invent an effective lie, he must design his falsehood under the guidance of that truth. On the other hand, a person who undertakes to bullshit his way through has much more freedom. His focus is panoramic rather than particular. He does not limit himself to inserting a certain falsehood at a specific point, and thus he is not constrained by the truths surrounding that point or intersecting it. (Frankfurt 2005, 51–52)

Therefore, it is not merely the fact that they do not have to respond *as competently* to their hedged bullshitting's challenges, it is that, by hedging, they can continue bullshitting in response in a way that is facile compared to the responses dictated by factive challenges to *unqualified* bullshit about the *same content*. The ease with which further bullshit follows hedged bullshit is markedly different when compared with its bare counterpart.

So, the hedging bullshitter has access to potent resources in response to challenges that the unqualified bullshitter lacks. But what happens when they are just wrong? After all, bullshitters have different priorities than accurately tracking reality.[39] Therefore, while their assertions will not *necessarily* be false, they are presumably going to

38 Is it *impossible* that the speaker continues to bullshit in response to (14b)? Probably not, but it is highly plausible that it is far more difficult to keep bullshitting in response to a challenge like (14b) than one like (13c)—as I am just about to discuss.

39 I am assuming Covaian bullshit but no matter what account of bullshit one endorses, truth-tracking does not take precedence.

be incorrect more often than non-bullshit assertions. And so, the frequent bullshitter will be regularly disseminating falsehoods, and thus will inevitably be known as an unreliable testifier, a consummate bullshitter, and no one will listen to them anyway—irrespective of their potential lack of susceptibility to challenges.

But what if hedging ameliorated *even these reputational worries*? Recall from §1.3 two normative effects of hedging:

> RESPONSIBILITY EFFECT
> Hedging makes a speaker less responsible for the content of their assertion.
>
> REPUTATION MANAGEMENT
> Hedging can exempt the hedged testimony from affecting the speaker's reputation.

Strikingly, these effects relate directly to and attenuate the possible reputation worries that stem from being (known as) an inveterate bullshitter. If the speaker's bullshit turns out to be false—say there were only 21,000 attendees at the rally, not 100,000 as was speculated—well, good job the bullshitter only said there *might* be six-figure attendance! The bullshitter in (2) did not proclaim to *know* that the truth was being withheld from us, so we cannot go overboard with opprobrium; if it turns out the speaker in (3) was completely wrong about the AI's anti-Caucasian slant, well, they were only speculating that they *they think may* know why.

By hedging their bullshit, speakers have vastly greater scope to bullshit more because their reputation is considerably more insulated. As van Elswyk puts it:

> [Hearers] either lose the ability to hold speakers accountable or have that ability significantly diminished. Since accountability is the recourse hearers have if a content turns out to be false or unknown by the speaker, hearers lack meaningful recourse for hedged testimony that is learned later to be defective. (van Elswyk 2023, 347)

This dovetails perfectly with the empirical data noted above, forming a sort of vicious cycle; Petrocelli (2018) notes that people are more reticent with bullshitting when they suspect they will be held accountable. As the van Elswyk quote above explains, however, hedging *specifically* reduces how much a speaker *can* be held accountable. Therefore, a hedging bullshitter can bullshit more because they will not be held as accountable, and because they know they will not be held as accountable, they will bullshit more. And this is all already built in to the picture of pragmatics and conversation we abide by every day, making these sorts of normative effects nigh on immutable in ordinary discourse.

Still, my opponent might respond that this all ultimately evens out because of REPUTATION MANAGEMENT. The idea is that while it is true that hedging protects one's reputation if they are ultimately wrong, hedging symmetrically insulates

against reputation *gain* if they are ultimately *right*.[40] So, the hedging bullshitter does indeed protect themselves (to some degree) if it turns out that they are wrong, but this ends up being a neutral move in the overarching language game because they will equally prevent similar reputation *gain* if they somehow happen to be right.

There are a couple of reasons this response is unconvincing. First, the bullshitter is presumably going to be wrong much more than right (recall affect takes precedence over accuracy), so the reputation gain missed out on is not even close to the reputation loss insulated against simply due to the ratio of incorrect to correct being hugely weighted in favour of the first. Second, a slightly more speculative point I call ASYMMETRIC REPUTATIONAL FORFEITURES. In essence, I want to outright reject this idea of proportional reputation stultification in both directions—*particularly* when it comes to the examples of bullshitting like (1)–(5) and similar cases. I maintain that the bullshitter has little to lose and lots to gain by hedging their bullshit. Specifically, I think their reputation *is* protected if (or, more likely, *when*) they are wrong, but I propose that if they are correct, they ultimately forego little or no reputation *gain*.

We can make this clearer by employing the examples of hedged bullshit. My thought is that the decision to hedge and not represent knowledge for the far-out possibilities like the World Health Organisation engaging in international assassination attempts, or six-figure attendance at political rallies *does* act as a very effective disclaimer[41] for preventing grand mal opprobrium when these statements are eventually shown to be incorrect. If, however, such varieties of low-probability bullshit somehow turned out *true*, I suspect the speakers would enjoy reputation gain indistinguishable from an unqualified assertion of the same bullshit. More concretely, suppose it came out that the WHO did indeed try to assassinate the Slovakian prime minister. Would the fact that the speaker of (1) hedged their assertion of this impinge upon the reputation spoils they would undoubtedly enjoy for 'predicting' it? My guess is not in the slightest. Of course, the WHO are obviously not organising political homicide, but (1)'s progenitor still has nigh on carte blanche to make this claim thanks to their hedge and suffer (at most) nominal reputation loss. Combine this with the fact that these bullshitters are making tens or hundreds of bullshit claims a day, it is presumably likely that they will *eventually* be correct, and thanks to ASYMMETRIC REPUTATIONAL FORFEITURES, can fully cash in on their reputational gains. Now, to do some hedging of my own, I imagine this is ultimately an empirical question of some sort which I do not have data on, but I think as some informed speculation, it passes the sniff test.

In sum, as has been discussed several times throughout, different challenges are licensed and different responses to those challenges permitted depending on whether

40 Again, compare with *hedging your bets*; hedging does not change expected value, it only reduces exposure.

41 As McCready notes: "Disclaimers act precisely like hedges in terms of their pragmatic function, by insulating communicators from potential negative consequences of error." (2014, 43).

an assertion is unqualified or hedged. However, when the hedged assertion is also a bullshit one, this results in significant differences in the responses the bullshitter is obliged to field to challenges. Hedging can play a dramatic role in making things much easier for the bullshitter: from easing the normative pressure they have to make a good response to a challenge of their bullshit, to opening the door for them to continue bullshitting with abandon. Moreover, hedging does a miraculous job in protecting the bullshitter's reputation, thus helping to ameliorate the classic reputational issues native to being a consummate bullshitter. Moreover, due to imbalance in their mainly saying falsities, they are greatly advantaged in hedging all their bullshit. And finally, should ASYMMETRIC REPUTATIONAL FORFEITURES be correct, then the bullshitter can reap all the advantages of hedging while suffering none of the disadvantages.[42]

1.6.2.3 A tension eased

§1.6.2.1 opened by outlining a tension at the heart of the bullshitting literature. I argued that (iii) of the triad can be eased by employing hedging in one's bullshit assertions as it allows the consummate bullshitter to bullshit more freely, with less risk of difficult challenge or reputation loss. Now, to be clear, I am not suggesting that hedging one's bullshit is a panacea, meaning one can bullshit as much as they want, forever, without repercussions. But its properties in alleviating the known, crucial weaknesses of unqualified bullshitting are arresting, and go a long way in mollifying this plausible discrepancy between (i), (ii), and (iii). The danger of hedged bullshitting is clear and we now can see why the bullshitters in (1)–(5) employed hedging—it is an essential tool in the deceiver's arsenal.

1.7 CONCLUDING REMARKS

ALETHIC ACCOUNTS of bullshit predict the impossibility of hedged bullshitting. We have seen it is not only possible but alarmingly actual. I have argued for two main theses in this chapter; first, that hedged bullshitting is a bona fide phenomenon and thus we ought to reject these standard theories and endorse Cova's competitor account; second, I argued that hedged bullshitting is not merely some niche occurrence to be wielded against the ALETHIC ACCOUNTS but a superlative form of deception that it is paramount we understand.

I called §1.6.2 *Getting away with bullshitting* but, by homing in on the details and mechanics of such hedged bullshitting, I hope to have gone some ways in starting to nullify such advantages hedging grants so that deceivers *cannot* get away with

42 It remains to be seen whether ASYMMETRIC REPUTATIONAL FORFEITURES applies to hedging in general, or only in the kinds of low-probability, bullshitting cases I have discussed. This is an interesting question that I think is ripe for further research.

it. Bullshit is, as we well know, rather dangerous. Peter van Elswyk's "Hedging in Discourse" opens with the pithy remark, "A speaker does not always come out and say it." (2024, 1) Now let me add: *neither does a bullshitter*.

Linking Interlude I

Towards the end of this opening chapter, I looked at *challenging* the hedged bullshitting in play. In lockstep with the hedging literature, I looked specifically at the challenges of *How do you know that?* or *You don't know that.* for the unqualified bullshit, and weaker varieties for the more dangerous *hedged* bullshit. But of course, these are not the *only* ways to challenge bullshit—or indeed *any* sorts of (what we take to be) falsehoods, misleading statements, and so on. Nor, we might think, are these the *best* ways. If the above challenges are more *questioning* the speaker and their assertions, we might think we are better off doing something *stronger* to properly tackle them. We might propose that we *argue* with the assertors, that we *object* to their assertions. The next chapter discusses this sort of objecting to asserted falsehoods. Specifically, it gives an answer (among other things) to the following question: what makes for an epistemically good objection?

Chapter 2

Objecting and quiescing, aptly

Objecting to asserted falsehood is an everyday practice. Despite its apparent ubiquity however, there remains a distinct dearth of discussion of objecting in the epistemology literature, with the small amount of extant scholarship focusing primarily on the prescriptive—viz., epistemic duties to object. The spotlight being placed so has elided a key area of research which seems to precede any prescriptive discussion; namely, evaluative norms about what makes for a good objection. After all, knowledge that one ought to object lacks utility if one does not also know how to well discharge such a duty. Thus, in this chapter, I build a complete picture of the evaluative normativity of objecting based on a Sosan performance-normative framework, neatly answering the question of what is an epistemically good objection—it is an apt one. The view is then extended to explain a novel area: the evaluative normativity of forbearance from objecting, what I call quiescing.

2.1 Introduction

We object to others' asserted falsehoods all the time. We frequently have our own asserted falsehoods objected to as well. Much of the time, one might think that, beyond it being merely good or helpful to object and correct someone's false belief, we in fact *ought* to object to such assertions. Suppose your friend wrongly states the time you are supposed to meet later. Not only would it be bad if you did not correct them (after all, you presumably *want* to spend time with your friend), one might propose that you in fact have a *duty* to object and correct their false belief about the time of rendezvous. Here, we would likely source such an ought in the *practical*—that is, you prudentially ought to object so that you can gain some practical end (in this case, successfully meeting up with your friend later).

In other scenarios, you might have a duty to object rooted in the *moral*. Perhaps someone asserts something sexist to you. Not only would it be bad if you did not object to such sexism (after all, you presumably think sexism is wrong), one might

contend that you *ought* to object and (at least try to) correct their false, sexist beliefs. This ought, intuitively, seems a moral one; perhaps we have a general moral duty to stand up to sexism, and, in this instance, it would be cashed out in the form of an objection to a sexist assertion. And while both cases are in some ways connected to or correlated with false beliefs, one could be forgiven for thinking that there is not much especially *epistemic* going on—particularly not as a source of any oughts or duties, anyway. Indeed, prima facie, it is rather difficult to construct a situation where there is (what looks like) epistemic normative pressure to object that does not carry with it superseding moral or practical import.[1] Difficult, perhaps, but not impossible:

> PUB
> It is 11:45pm on a Saturday. Alex and Bella are at the pub. Bella would quite like to go home, but nevertheless asks Alex if he would like to get another drink. Alex says that he would prefer to just leave, but adds that it is a moot point anyway as the pub closes at midnight. Alex is actually incorrect—the pub does not close at midnight on Saturdays—and Bella knows this. She considers staying quiet but feels that she should not do so, and thus objects and corrects Alex's false belief. She points to a sign above the bar saying, "Different closing hours at weekends", and explains that while the pub does close at midnight during the week, it is open until 1am on Saturdays and Sundays.

Assuming that Bella is not wildly irrational for feeling like she ought to object (or ought not to stay quiet), where would we locate this ought? We can immediately discard the prudential—both Alex and Bella want to leave, so nothing of practical value would be lost should Bella remain quiet (in fact, to meet her practical goal of going home, we might even think she *prudentially* ought to remain quiet). The moral case seems rather weak; we could perhaps claim that Bella would be deceptive or lying via omission should she not object, and that this is ethically problematic, and thus she *morally* ought to object. However, the case at hand is such low stakes that I do not find there to be much of an intuitive pull towards this line of thinking. Moreover if we bought into this idea, then we would have a *moral* duty to correct anyone's (potentially) wrong information on the most trivial and inconsequential matters, and—even setting aside potential demandingness worries—it is just not clear what is distinctly *moral* about this. In fact, it seems to me that the *epistemic* provides the simplest and most sensible explanation; it is straightforwardly epistemically bad to allow someone to maintain a false belief when it appears that you could object and correct it. As Jennifer Lackey has recently described: "If it is in our power to prevent something epistemically bad from happening through very little effort on our part, we ought,

1 See Johnson (2018, 121–124) and Lackey (2020b, 37–39) for the original cases and argumentation where we remove moral and/or prudential import, leaving behind only *epistemic* normative pressure. I follow them here with a case of my own.

epistemically, to do it." (2020a, 287) If you accept this diagnosis, then, in some circumstances, we have plainly *epistemic* duties.

In this chapter, I take seriously this idea of positive epistemic oughts, with *objecting* taking centre stage.[2] I am not the first to do this; Casey Johnson (2018) and the aforementioned Jennifer Lackey (2020b) both recently argued for a distinctly *epistemic* duty to object.[3] Nevertheless, it is fair to say that, despite the apparent ubiquity of the actual practice of objecting to falsehoods, there is a definite dearth of discussion of the phenomenon in the (social) epistemology literature, with the small amount of extant scholarship focusing primarily on the *prescriptive*—that is, if an epistemic duty to object is plausible, when do agents have such a duty, what does such a duty look like, and so on. The spotlight being placed so, however, has elided a key area of research that, to my mind, comes before any prescriptive discussion: namely, *evaluative* norms about what makes for an epistemically good objection. After all, knowing *when* one ought to object seems to lack utility if one does not also carry with them an idea of how to discharge the duty effectively. Suppose through my moral education I know that I have a moral duty to help those collapsed in the street, but when I eventually come across someone in such a state, I act to fulfil the duty by running over to them and incessantly shouting at them, "are you okay?!" for ten minutes instead of doing anything actually helpful. The fact that I knew I had a duty to help proved rather insignificant in the end because I had no idea how to execute the act of *helping those collapsed on the street*. In the same vein then, being *aware* that one has a distinctly epistemic duty to object is not of much use if one is *unaware* of what makes for an epistemically good objection.

That is the main contribution of this chapter: drawing on a Sosan performance-normative framework (Sosa 2007, 2015, 2021), I build a comprehensive picture of the evaluative normativity of objecting, elegantly answering this question of the structure of epistemically good objections. I then demonstrate that the framework possesses the resources to also explain the evaluative normativity of the novel area of *forbearing* from objecting—what I call *quiescing*. One discussion precedes the above, however: the metaphysical question of "what is an objection?" It is there, then, that we first turn.

2 The foregoing was not intended as a slam-dunk argument to 'prove' the existence of an epistemic duty to object (nor a novel one—see the previous note 1), but rather as a helpful introduction to the topic at hand, so one might reasonably disagree or find such a positive epistemic duty dubious. Nevertheless, I will walk in lockstep with the extant literature and grant this duty—and in this way of viewing things, even sceptics should take note: after all, a true conditional can have a false antecedent.

3 See Lackey (2020a, 2021) and Terzian and Corbalán (2021) for essentially the remaining extant work on epistemic objecting oughts specifically. For further ideas in the vicinity (although they do not use the same terminology as the aforementioned authors), see Cassam (2019b, chap. 5) and Mill (1859, chap. 2).

2.2 WHAT IS AN OBJECTION?

In her work on objecting, Lackey has at different times proposed two different accounts of objections.[4] The more recent one is unsurprisingly intended to be an improvement on its predecessor, although I actually think it is arguably less plausible. Let us start with this latest one:

> EVIDENCE ACCOUNT
> "Objections [are] assertions that are added to a conversational context with the aim of adding to the communal pool of evidence."[5] (Lackey 2020a, 287)

So, if an interlocutor were to assert that p, and you take p to be false, you might object. This would entail, say, offering up some evidence or argument that suggests $\neg p$. Assuming that meeting the aim is how we dictate success, then the main problem facing the EVIDENCE ACCOUNT quickly becomes transparent; it is simply too liberal, countenancing far too many cases of clearly *unsuccessful* objections as *successful*. Recall PUB from above, and suppose that after Bella objects, explaining that the pub is open until 1:00am on Saturdays, Alex is certain he is correct and so does not heed her at all, maintaining his false belief of a midnight closing time. Despite the objection's content not being taken up by Alex, Bella *has* still added to the communal pool of evidence here (she spoke up, presented facts, and so on). Therefore, according to the EVIDENCE ACCOUNT, she meets the aim and thus achieves a *successful* objection. This is a clear misdiagnosis, however. If Alex entirely disregards her response and holds onto his false belief in the face of Bella's (accurate, cogent) objection, it would be very odd to consider this a success. In fact, this seems like a paradigmatic example of an *unsuccessful* objection—irrespective of the fact that Bella has presumably done as well she could. So, merely *adding to the communal pool of evidence* as the success condition is evidently too weak, and thus the EVIDENCE ACCOUNT does not make for an adequate account of objections.

Lackey's earlier definition (or at least a reading of it therein) has, I think, more promise. It is as follows:

> RECORD ACCOUNT
> "Objections [are] assertions that are added to a conversational context with the aim of correcting the record." (Lackey 2020b, 36)

4 One might wonder what Johnson (2018) took an objection to be considering that I just described her as the other main player in this literature. In fact, she focuses on the more deflationary "voicing disagreement" so as, I take it, to more cleanly home in on *epistemic obligations* themselves and her excellent arguments for them. Nevertheless, Lackey *does* explicitly define objections, and having a plausible account of objections is a crucial part of my chapter to come, hence this section.

5 To be clear, I am following Lackey (and the rest of the literature) in taking "objections" as responses to asserted propositions. One might "object to capitalism" in a general sense, for instance, but that is quite a different phenomenon from the one I am interested in for this chapter. Cf. note 6.

We can consider two different readings of how one might meet the aim of "correcting the record":

> TB-RECORD ACCOUNT
> Target takes up the relevant true belief(s).
>
> ¬FB-RECORD ACCOUNT
> Target jettisons the relevant false belief(s).

I will propose that the second reading is most plausible. First, however, one might wonder if this "correcting the record" has anything to do with beliefs or indeed success conditions and is instead simply referring to tally marks on *conversational* records or ledgers. This may very well be the case and be what Lackey intended—her exact intentions are left unspecified with the attention mainly being centred on discharging the duty of objecting not requiring "the likelihood of acceptance," (i.e., success) and merely necessitating an "intention to have your dissenting voice noted." (2020b, 36) Of course, I agree that objecting *simpliciter* certainly does not require success, just as shooting at goal does not require a goal to be considered *shooting at goal*. Nevertheless, objecting *successfully* (i.e., correcting the record, acceptance, whatever this precisely entails) trivially *does* at the very least require *success* and the objecting literature is replete with reference to and discussion of (false) beliefs and the changing of minds (see any of the literature cited so far, or e.g., McCormick (2023), Battaly (2021), and McIntyre (2021)). Thus, these readings of the RECORD ACCOUNT, and the DOXASTIC ACCOUNT I soon land on, are angles that are both commensurate with and supported by the extant literature—and, so far as I can tell, my DOXASTIC ACCOUNT makes for the first definition of objecting actually argued for as opposed to merely assumed or granted.

Let us not get ahead of ourselves, however, and first take a look at this initial interpretation of the RECORD ACCOUNT, the TB-RECORD ACCOUNT. While a target taking up the relevant true belief is *sufficient* for a successful objection, it is not necessary, meaning the account proves slightly too conservative to make for a fully adequate account of objecting. Consider an alternate PUB case:

> PUB*
> It is 11:45pm on a Saturday. Alex and Bella are at the pub. Bella asks Alex if he wants to get another drink but Alex replies that there is no point as the pub closes in fifteen minutes. Alex is incorrect—the pub closes at 1:00am on weekends—and Bella knows this, so she objects. She points to a sign above the bar saying, "Different closing hours at weekends", and explains that while the pub does close at midnight during the week, it is open until later during the weekend. Alex acquiesces and stops believing that the pub closes at midnight on the weekends.

In this version of PUB, Alex does not take up the relevant true belief (that the pub closes at 1:00am), all he has done is jettison his prior false belief (that the pub closes at midnight). He is essentially suspended on the question of "what time does the pub close?"—although he has removed as possible answers any time up to and including midnight. According to the TB-RECORD ACCOUNT, this objection fails to meet the success condition of the target taking up the relevant true belief and so is *unsuccessful*. This seems mistaken, however. Intuitively, Bella's objection looks successful insofar as Alex no longer holds the relevant false belief—promoting him to holding a further relevant *true belief* looks auxiliary. In fact, I suspect the slippage here is that objections are generally accompanied by an additional assertion that results in the target taking up the relevant true belief once the relevant false belief has been jettisoned (or the objection simply contains the true belief in some way). A final variant of PUB makes this clear:

> PUB**
> It is 11:45pm on a Saturday. Alex and Bella are at the pub. Bella asks Alex if he wants to get another drink but Alex replies that there is no point as the pub closes in fifteen minutes. Alex is incorrect—the pub closes at 1:00am on weekends—and Bella knows this, so she objects. She points to a sign above the bar saying, "Different closing hours at weekends", and explains that while the pub does close at midnight during the week, it is open until later during the weekend. Alex acquiesces and stops believing that the pub closes at midnight on the weekends. Bella then tells him that the actual closing time is 1:00am. Alex subsequently takes up the relevant true belief that the pub closes at 1:00am on Saturdays.

Here, the record has indeed been corrected to a true belief uptake in Alex but it is clear that *objection* characterisation applies only to the first part where she responds to his initial mistaken assertion and provides evidence against it. The new true belief stemmed from Bella's following assertion about the exact closing time which was clearly discrete from the objection itself.

Pulling all this together, we can infer that the TB-RECORD ACCOUNT is slightly too exclusionary, and something in the region of the ¬FB-RECORD ACCOUNT and its focus on the abdication of falsity would be more accurate. Therefore, here is the account of objections we will proceed with for this chapter:

> DOXASTIC ACCOUNT
> An objection is a response to an interlocutor's false assertion that aims at the interlocutor's discarding of the relevant false belief.[6]

6 To be clear, I am taking this aim to be a *constitutive* one—viz., part of what an objection *is* is a response that aims at the relevant false belief's discarding by the interlocutor. What this mainly serves to do here is that if one speaks up about something merely to look cool in front of their friends or to win

From this definition, we can make explicit how exactly an objection would be successful as well:

> SUCCESS CONDITION
> A successful objection is one that results in the interlocutor discarding their relevant false belief.

We can see that this account correctly diagnoses the cases given so far; the version where Alex maintains his false belief comes out as unsuccessful due to the discarding falsity requirement; the variant PUB* where Alex simply jettisons his false belief without taking up the relevant true belief is considered successful; *and* the rendition of PUB** where he does further believe the correct closing time is also cashed out as meeting the SUCCESS CONDITION as it likewise involves the abandonment of the false closing time belief. Plausibly, then, a satisfactory answer to the question of *What is an objection?* has been located in the DOXASTIC ACCOUNT and so we are ready to turn to the main component of the chapter, developing evaluative norms that enable us to answer the normative question of *What is an epistemically good objection?*

2.3 A SIMPLE SOLUTION

Given the DOXASTIC ACCOUNT and its derived SUCCESS CONDITION above, we might think there is, staring us in the face, a very simple solution to this evaluative question regarding epistemically good objections. We found that *successful* objections were those that resulted in one's interlocutor discarding their relevant false belief, and so we might just endorse a sort of identity claim: a good objection *just is* a successful objection. In other words, we could propose an evaluative norm like the following:

> SIMPLE NORM
> An objection is epistemically good iff it meets the SUCCESS CONDITION.

Indeed, both our successful objections of PUB* and PUB** certainly look epistemically good (Bella made accurate and cogent points and presented truth-conducive evidence resulting in a false belief discarding from Alex) which is presumably a good start for the SIMPLE NORM. Nevertheless, we will now see that the SIMPLE NORM quickly generates unintuitive results once the cases get slightly more complicated. Consider this first case:

> UNSUCCESSFUL SAMARITAN
> Andrew has somehow stumbled into a convention of Holocaust deniers. Someone confronts him and starts talking about the Holocaust being a

a bet, then even if this syntactically appeared like an objection, it would not count as one as it is not aiming at the discarding of a false belief. For classic work on constitutive aims related to assertion and games, see Williamson (2000, chap. 11).

> hoax. Andrew knows this is false and so objects. He presents good evidence, argues well, and generally does an admirable job of proving the reality of the Holocaust. Unfortunately, his interlocutor does not change their beliefs at all, and indeed argues back, rejecting his evidence, and calling him a liar and a fool.

According to the SIMPLE NORM then, this objection is not a good one; it does not meet the SUCCESS CONDITION and so cannot be considered epistemically good. Of course, there is a sense in which this is correct—Andrew's objection clearly was not successful in changing any minds, nor was it successful per the exact terms of the SUCCESS CONDITION. Nevertheless, we might think this is not grounds to simultaneously derive its *complete* lack of goodness. In fact, it strikes me that there is a clear intuition that Andrew did something well here. This is not just a moral pull either.[7] Pace the SIMPLE NORM, Andrew's objection was epistemically good in a myriad of ways: it was well evidenced, contained good argumentation, proved his conclusion, and so on. More technically, we might say Andrew's assertions were knowledgable, his testimony reliable, his steadfastness in the face of disagreement virtuous. A straightforward evaluation of *not epistemically good because unsuccessful* from the SIMPLE NORM misses important dimensions of Andrew's objection, and casts doubt on the sufficiency of the SIMPLE NORM as a fitting evaluative norm for epistemically good objections.

Nevertheless, it would be quick to jettison this solution already. Maybe, with a little work, we could explain away the apparent epistemic positives of Andrew's objection and rescue the *not epistemically good* judgement. Perhaps the lack of success overwhelms any auxiliary epistemic goodness. Possibly. However, UNSUCCESSFUL SAMARITAN is not unique in its having a rather unintuitive diagnosis from our simple norm. In fact, the verdicts only get worse. Consider a different vignette where the elements are somewhat reversed from the previous case:

> ATLANTIS ATTESTATION
> Bryony is talking with climate change sceptic, Charlie. Charlie asserts that climate change is nothing to worry about, and nor is it even really happening. Bryony is aware that this is not true, and so objects. She does not know much about climate science, however, and is not sure how best to support her case. She decides some creative license is her best bet and cites the supposed fact that the wondrous city of Atlantis was recently drowned due to rising sea levels which were caused by climate change. Upon hearing of this horrific disaster, Charlie immediately comes round, and changes his outlook on climate change.

7 That is, one may suppose that Andrew succeeded here merely by standing up and objecting to a Holocaust denier, irrespective of whether any beliefs are revised. While this is correct in a certain sense, it speaks to a sort of *moral* success, and not the *epistemic* success that is of central concern in this chapter.

This objection is epistemically good, says the SIMPLE NORM, because it clearly meets the SUCCESS CONDITION. Evidently, however, it is epistemically bad to say untrue things, to feed people false information, to lie,[8] even if it ultimately has a good end—viz., getting Charlie to jettison his false belief about climate change. The details of this objection might make for good examples of *sophistry*, or *persuasion*, which are certainly related areas to the *objecting* taking point here, but that does not make Bryony's response an *epistemically good objection*. One might even argue Bryony's actions here are all things considered 'worth it' but that is entirely commensurate with having a more fine-grained epistemic evaluation than simply "good". In short, it seems that an unambiguous judgement of *epistemically good because successful* from the SIMPLE NORM would miss out on crucial elements of Bryony's objection that should adulterate such an evaluation.

So far we have seen the SIMPLE NORM give fairly naive evaluations of cases of both failure and success that look unsatisfactory and oversimplified. This would be enough, I think, to set aside the norm and look elsewhere, but let us consider a final case anyway, in the name of rigour:

> OPPORTUNE OBJECTION
> Diana's acquaintance, Ewen, tells her of his belief that vaccines are slow-fuse death sentences. Diana objects as she is certain this is incorrect. She offers up a plethora of solid evidence showing otherwise, explaining that vaccines are one of humankind's greatest inventions and justifying this claim in full. Ewen is just about to dismiss Diana's claims as the government propaganda that he has read so much about when a tulip petal blows into his mouth, briefly choking him. This reminds him of his dear grandmother who so loves tulips, and always advised him to be open-minded to genuine people. In a flash of inspiration, Ewen decides to listen to Diana, and finds his false belief about vaccines dissipates.

This vignette is, in a strange way, simpler *and* more complex than the previous two. On the one hand, the SIMPLE NORM's positive epistemic evaluation does not prima facie appear as ill-fitting as the prior judgements did—after all, Ewen does jettison the relevant false belief *and* Diana's objection is epistemically top-class in the way we want objections to be. That is, it is accurately evidenced, well argued, and wholly justified. And yet, on the other hand, there is the lingering notion that *something* is slightly amiss. Is not the fact that Ewen *nearly* instantly rejected Diana's objection wholesale before, through apparent happenstance, a tenuously connected memory was evoked which lead him to suddenly listen enough to somewhat sully the pure *epistemically good* evaluation? We might think, at the very least, some problematic luck

8 Of course, these sorts of "noble lies" are morally bad too—perhaps more morally problematic than epistemically. That still does not mean they are *not* epistemically bad, however, even if we might usually focus more on the ethical side.

is at play here, and it would be quite an asset for our evaluative norm(s) if they could offer up an explanation of what is happening in OPPORTUNE OBJECTION. Whatever the exact diagnosis of the case ought to be, for present purposes, it is apparent that the SIMPLE NORM does not have anything to offer in the way of such an explanation, making this case the final—and fatal—mark against its card.

To sum up, in a potentially unnecessary show of force, the prosaic SIMPLE NORM is unsurprisingly too simple and transparently untenable as a solution to our guiding question of what makes for an epistemically good objection. The key takeaway from the three cases is that, while success undoubtedly has a major role to play in our epistemic evaluation of objections, there is evidently far more to a response to falsity's epistemic goodness than the mere satisfaction of the SUCCESS CONDITION. Despite the fact that one might have suspected from the off that the SIMPLE NORM's prospects were on the lower side however, our time here in this section was not wasted; we now have the three above vignettes in addition to our PUB cases from before which we can use to test the extensional adequacy of any account we put forward. The next step of this chapter, then, is to develop my performance-normative account of the evaluative normativity of objecting, and see how it fares.

2.4 INTRODUCING PERFORMANCE NORMATIVITY

The performance-normative framework has been the pivotal tool in the employ of Ernest Sosa (2007, 2015, 2021) in his project of analysing knowledge—what he calls *gnoseology*. This project is (perhaps along with archery) the most well-known use of performance normativity that most (epistemologists, at least) will be familiar with. Therefore, it is through this virtue-theoretic, epistemic lens that I will do the majority of the exegesis on the framework throughout this chapter. Nevertheless, it is crucial to keep in mind that the performance-normative framework is not exhausted by Sosan reliabilist virtue epistemology. Rather, performance normativity is just a schema for evaluating *any* performance, be that believing, figure skating, scoring three-pointers, opening a window, or perhaps even objecting. The fact that believing is the pre-eminent use of the framework in philosophy is essentially a contingent historical fact because Sosa is an epistemologist who came up with the framework and then used it to analyse knowledge. It is important to explicitly flag this and ensure we do not confuse the two as a single entity.[9]

With these provisos in place, we can now briefly explain Sosa's belief-centric performance normativity as a helpful start to understand the more general phenomenon. The whole virtue epistemology picture is driven by the single fundamental idea that belief aims at truth.[10] From this aim, a *success condition* of beliefs is derived—that is, a

9 Some are even sceptical that beliefs aim at anything so would contend that this belief-centric use is inappropriate. See, for example, a recent paper from Vermaire (2024).

10 To be clear, I am setting aside *Judgment and Agency* (Sosa 2015) where Sosa distinguishes between

belief is successful when it is true. *Success* is one central dimension through which it can be evaluated, so too is *competence*. This is the idea that, irrespective of *whether* the belief is true or not, we can evaluate *how* it was formed, and whether it was formed *competently*—that is, in such a way that reliably arrives at truth. The final, vital dimension of evaluation is that of *aptness*: a performance is apt when it is successful through competence—a belief is apt when it is true *because* it was competently formed.[11] For Sosa, apt belief is *knowledge*.[12]

That is a basic overview of Sosa's virtue epistemology. Notice again there is nothing *inherently* belief-y about the *success/competence/aptness* framework and everything said can easily apply *mutatis mutandis* to any other performance with an internal aim. Many scholars have in fact recently shown as much with other areas of philosophy such as trust (Carter 2020, 2022, 2023b, 2024), blame (Simion 2021), basing (Titus and Carter 2024), questions (Carter and Willard-Kyle, forthcoming), and deference (Carter and Kallestrup, forthcoming). For instance, when it comes to trust, according to Carter, the aim of trusting is that the trusted will take care of things as entrusted by the trustee, and then from this we can generate a full performance-normative picture. In principle, so long as an internal aim exists, this schema can be used to evaluate *any* performance type. The implication here is not a subtle one; we have in fact already arrived at a SUCCESS CONDITION for objections, so we are essentially a third of the way there towards generating our performance-normative, evaluative norms for objecting.

Before we explicitly lay out our new norms, however, let us first pre-emptively head off an objection. I already noted that we do not need to posit some deep metaphysical connection between *believing* and *objecting* for the latter to be an appropriate fit for the performance-normative framework—just like Sosa does not posit a deep metaphysical connection between, say, *tennis* (Sosa 2021, 166) and *believing* when discussing their respective performance-normative evaluations. Indeed, to reiterate once more, the only commonality required to run through all these discrete performances is that they have an internal aim. Belief aims at truth, archery aims at hitting the target, snooker aims at potting the ball, and so on. From these rocks, performance-normative churches can be built. Thus, for objecting to be apt for a performance-normative treatment, it must similarly possess such an internal aim. This is where my opponent might be sceptical that objecting really fits in and claim that objections have no such goal.

Nevertheless, I do not think this is a worry we need to take seriously. Recall in

judgemental belief and alethic affirmation to focus on the classic "belief aims at truth" as this simpler version is more than adequate for our explanatory purposes here.

11 There is a further level of evaluation called "full aptness", that is, "aptness on the first order guided by apt awareness on the second order that the first order performance would be apt." (85) Again, this extra level of complexity is not required for our exegetical purposes here, so I will side-step it.

12 Or, at least, animal knowledge. Sosa bifurcates knowledge between animal and reflective—for an excellent overview of all this see the recent Carter (2023a).

§2.2, where we arrived at our DOXASTIC ACCOUNT with its internal aim of eliminating the relevant false belief in the interlocutor. This was not some arbitrary choice of language so as to shoehorn in Sosa now. In fact, as we saw, both of Lackey's different accounts of objections (the only definitions on the market in this small literature), the first of which was the precursor to our DOXASTIC ACCOUNT, already explicitly couched objecting in terms of *aims*—and Lackey certainly was not attempting anything Sosan in her work on the (epistemic) duty to object.[13]

I am not just blindly following Lackey in her characterisation of objecting having an aim either. It is, I think, the entirely natural and plausible way of thinking about our practice of objecting. Take the example with which I opened this chapter about the friend stating the wrong rendezvous time; it is not clear how we could even make sense of *why* we object to (what we take to be) falsities if it is not to try to correct the relevant false belief in the assertor. Indeed, when one objects to their friend about the rendezvous time, the seminar location, or the capital city of Australia, it would be rather baffling to interpret this as, for instance, that person signalling their membership (Tappin, Pennycook, and Rand 2021) of the *Canberra is the capital city* group, while their friend resolutely sits in *Sydney*.[14] Instead, the intuitive analysis is that the person knows their friend has a false belief, they do not want them to have that false belief any more, thus they object to try to allay them of this untruth they possess. Their objection succeeds if and only if their friend stops believing Sydney is the capital city.[15]

We are now in a position to formally introduce our SUCCESS NORM, the first of the three evaluative norms we will build for objecting from the performance-normative schema:

> SUCCESS NORM
> An objection is successful iff it results in the interlocutor discarding their relevant false belief.

Note that this is just a recasting of the SUCCESS CONDITION but in biconditional form

13 Indeed, the Sosan framework is entirely silent on prescriptive normativity and works through conditional prescriptions dependent on intentional aims. We will see this later when we discuss *suspending* in §2.7.

14 Of course, *sometimes* we might speak up about something merely to show that we disagree, to signal group membership, or protect some identity of ours, but, importantly, this is not *objecting* as I am conceiving of it here—despite the fact that it might look quite similar. Recall from note 6, that the aim in the DOXASTIC ACCOUNT is a constitutive one, so any kind of speaking up that does not aim at false belief revision is not an objection. Perhaps the examples above are cases of *flagging disagreement* that come with their own aims and success conditions but none of this is ultimately relevant to the project at hand. I will briefly look at *flagging disagreement* vs. *objecting* again in §2.7.

15 There is potentially an elephant in the room here regarding the *philosophical* objection familiar from Q&A sessions following talks. I would like to remain agnostic on this. I am tempted to say that the philosophical objection is a more technical use of the word objection that is different from the sort I am interested in here, thus we should not be led astray by any polysemy. On the other hand, it does seem like a lot of philosophical objections plausibly aim at the discarding of falsity. Either way, for simplicity's sake, I will assume I am not saying anything about philosophical objections in this chapter.

to more accurately follow the Sosan way of doing evaluative norms. Note further that if we stopped here and this were our *only* norm then we would essentially be back to where we were at the beginning of §2.3 with the SIMPLE NORM. Except, of course, that is manifestly *not* what the performance normativity framework entails—we have two more norms to develop before the picture is complete. The APTNESS NORM is facile to generate insofar as it is built upon the SUCCESS NORM and the COMPETENCE NORM—successful *because* competent—and while we are now in possession of the former of those two elements, we are obviously still lacking the latter. To competences, then, we will now go.

2.5 COMPETENCES

2.5.1 The general picture

Competences are dispositions to perform well in some area.[16] There is no *one-size-fits-all* account of "performing well"—different domains determine different declarations of a good or reliable performance. For example, a professional baseball player hitting the ball 25–30% of the time would constitute a *baseball-playing competence*, while only potting 25–30% of the balls you attempt in snooker would not even approximate a *snooker-playing competence*.[17]

Just as the performance normativity framework breaks down into three constituent parts (*success/competence/aptness*), so too do *competences*; SEAT: internal disposition to succeed; SHAPE: internal condition of the agent; and SITUATION: external factors. This in turn generates three different levels of disposition: *innermost competence* (SEAT), *inner competence* (SEAT *and* SHAPE), and *complete competence* (SEAT, SHAPE, *and* SITUATION). This is all rather theoretically laden, so consider the classic illuminating example from Sosa: one's car-driving competence. At the first level, SEAT, if one possesses the SEAT of a car-driving competence, then one has the general ability to drive a car—viz., a disposition to succeed were they to try to drive. This is an *innermost competence* that is retained even if one is synchronically unable to manifest the ability—say, because one is asleep or tied up. This leads nicely into the second level, SHAPE, which refers to the agent's current internal make-up. In order to manifest an *inner competence*, not only must one have the requisite ability (the SEAT), one must also be in the correct SHAPE—that is, sober, awake, untied, and so on. Finally, to achieve a *complete competence* for driving, the prior two levels must be properly in place, and the environment (the SITUATION) must be suitably conducive to a successful manifestation. For driving, the SITUATION must be such that the tyres have sufficient air

16 The following exegesis drawn primarily from Sosa (2010) but see also Sosa (2017, chap. 12). One difference is what Sosa calls in the earlier work "SEAT" he later calls "skill". This does not affect anything here so I will stick with the former terminology.

17 Professional snooker players tend to have a 90%+ pot success rate in professional matches.

pressure, the road is not waterlogged or especially icy, visibility is adequate, and so on.

There is one final piece of theoretical machinery in need of elucidation: the *trigger-manifestation conditional*. We employ this *trigger-manifestation conditional* to test for the general ability which corresponds to the SEAT—that is, the internal disposition to succeed. Consider, for instance, Sosa's archer. For an archer to have the SEAT of archery, the general disposition to hit the target with an arrow, it must be the case that the following *trigger-manifestation conditional* holds: if the archer were to shoot, they would likely hit the target. If this *trigger-manifestation conditional* comes out as true about someone, then they have the SEAT of archery—in other words, they have the disposition to succeed qua archer.

Nevertheless, while this is true in a vacuum, clearly other factors are also salient in whether the *trigger-manifestation conditional* holds; the archer must not be hallucinating phantom targets, say; the fletching on the arrows must not be damaged, presumably. Where have we seen such factors before? Of course, in the SHAPE and SITUATION. This set-up generalises to all performances: a skill or ability, a SEAT, is identified if the *trigger-manifestation conditional* holds, which in turn will require proper SHAPE and appropriate SITUATION. If all these aspects—the conditional, the internal factors, the external realities—hold appropriately, then we arrive at a *complete competence*. In order to meet a competence norm (and thus an aptness norm) one must manifest the requisite complete competence relating to whatever performance is in play.

2.5.2 An objecting competence

What might an *objecting complete competence* look like? More specifically, how would we precisify first the SEAT/*trigger-manifestation conditional* for the internal disposition to successfully object, second the appropriate SHAPE necessary for this, and finally the proper SITUATION such that a *complete competence* can be manifested by an objector? The first is not particularly complicated thanks to the work already done so far in this chapter; a simple conditional gets the work done:

> *Objecting trigger-manifestation conditional*
> If one were to object, one would likely object successfully.

Note that first this automatically brings in the necessary prerequisites; as we have been talking about objecting since §2.2, it requires an initial falsity to respond to, and if one is not aiming at an interlocutor's discarding of their relevant false belief, one is not objecting. Thus, the use of "object" in the antecedent of the *trigger-manifestation conditional* already tells us that an objector is aiming at the elimination of a false belief. A second key point to note is this *trigger-manifestation conditional*, much like a *believing* variant, is one that will come out true for the vast majority of agents (granted proper

SHAPE and appropriate SITUATION, of course, which will be discussed shortly). This is not a bullet to bite, as I have mentioned several times, small and trivial objections to interlocutors about seminar locations and shop closing times happen frequently, and most sophisticated enough agents are capable of allaying such false beliefs—one does not need to have studied for four years at objecting school to stop one from falsely believing that classes start on the 23rd and not the 24th.

The exact ability embedded in such a SEAT generated from this *trigger-manifestation conditional* is likely everyday skills to do with presenting countervailing evidence, reasonable argument, awareness of simple inference, and so on—essentially what people do every time they correct someone. Nevertheless, much as Sosa does not talk about his archer's elbow height or the timing of their breathing when drawing the arrow, nor do I need to say much about the specific ability here, it is enough that the *trigger-manifestation conditional* holds and grants *some* general disposition to succeed.

Turning now to SHAPE, it is also initially straightforward insofar as the proper internal condition for objecting is going to be along similar lines as those for most performances—that is, a lack of whatever internal conditions that would inhibit the performance. More specifically, we could say that one must be cognitively functioning above some threshold such that they are awake, somewhat circumspect, not blind drunk. The question of whether there are *objecting-specific* SHAPE features is more interesting—and less clear. It seems plausible that a prospective objector should possess some relevant knowledge about the topic at hand. For instance, only knowing *that* the assertion is false is likely not enough to object successfully, one probably would need at least a little more detail relating to *why* it is false, or how one knows it is false. Even in our trivial cases about seminar locations or shop closing times, a mere response of "wrong!" (even assuming this is an accurate call) is unlikely to prove sufficient, presenting at least a small amount of further evidence or knowledge about the matter at hand looks necessary to achieve a successful objection. So, we can say that, along with standard cognitive function, knowledge of the subject at hand is required for proper SHAPE.

Finally, let us consider SITUATION. When Sosa discusses the relevant external factors for his archer or his believer, he is primarily concerned with *physical* factors of the environment—gusts of wind or fake barns— and while they are certainly somewhat salient here insofar as we likely would not test for a complete objecting competence mid-hurricane or in the vacuum of space, the far more interesting (and pertinent) situational elements for objections are *social*. To start with a simple example, it would not count against one's objecting competence were they to fail in meeting the SUCCESS NORM during a debate competition where their opponent is specifically trying to respond in kind and not be convinced[18]—this scenario is clearly antagonistic to belief

18 For the sake of argument, let us also assume that the objector is indeed aiming at false belief revision in their interlocutor and not just trying to score well in the competition.

revision. The moral here is a familiar one irrespective of the exact performance: if the circumstances at hand are extraordinary making SITUATION essentially antithetical to the *inner competence* (SEAT and SHAPE) on display, then of course it has no bearing on one's complete competence when properly situated.[19] This is not to say either that the situation must be such that interlocutor belief revision is *facile*, instead it is merely that polarised, charged environments where anyone discarding a false belief would nigh on be a miracle are not what make up an appropriate objecting SITUATION. Indeed, this would involve making the unprincipled move of setting aside all the minor and trivial objections about matters of little consequence we do every day and taking the standard case as being objecting to a Nazi about their false and reprehensible views. This would also presumably have the knock-on effect of meaning that one's complete driving competence could now be tested for on ice rinks. These are clearly defective diagnoses.

Let us take stock. In this section, we built a *complete objecting competence*. At the first level, we have the SEAT, the internal disposition to object successfully, which corresponds to the following *trigger-manifestation conditional*: If one were to object, one would likely object successfully. The details of the antecedent and the consequent are filled in by the DOXASTIC ACCOUNT and the SUCCESS CONDITION/SUCCESS NORM, respectively. This conditional's holding relies on the objector being in proper SHAPE and appropriate SITUATION. The former we cashed out in terms of sound cognitive function and subject knowledge while the latter referred primarily to social circumstances surrounding the target of the objection—viz., an environment not wholly inimical to false belief revision. If these three elements are suitably in place for an objector, they can be said to have a *complete objecting competence*, giving us the following COMPETENCE NORM:

> COMPETENCE NORM
> An objection is competent iff its attestation manifests a *complete objecting competence*.

With part one and two of the triad now complete, we can also lay out our APTNESS NORM:

> APTNESS NORM
> An objection is apt iff its success manifests a complete objecting competence.

With our three, performance-normative, evaluative norms, we have completed our picture of the evaluative normativity of objecting. Recall our guiding question—*What*

19 To be clear, the SITUATION must be extraordinary-cum-antithetical *indexed* to the performance at hand. It may well be the case that the external environment for deep-sea welding is essentially antithetical to the performance of welding, but that is not *extraordinary* for that specific performance—indeed, it is the standard case. Thus, this would not count as an abnormal SITUATION for a deep-sea welder and we *would* test them for a *complete competence* in that high-pressure environment.

is an epistemically good objection?—that we have been on this quest to answer. Well, just as *apt belief* is *knowledge*, the "gold standard" of belief, so too then would an *apt objection* be the gold standard of *objections*, meaning our central question has a simple answer: an epistemically good objection is an apt one. Prima facie, this is an elegant solution. But now we must put it to the test and see how it fares with the vignettes that plagued the SIMPLE NORM.

2.6 TESTING FOR EXTENSIONAL ADEQUACY

Recall that in §2.3, the SIMPLE NORM was unable to capture the complexities at play in the three vignettes, and gave overly naive and simplistic judgements about whether the cases were epistemically good or not. What we have now is a more detailed and sophisticated picture of the evaluative normativity of objecting, meaning, for each case, we can test the protagonist's objecting along *three* different dimensions—namely, the SUCCESS NORM, the COMPETENCE NORM, and the APTNESS NORM. If all goes well then all the intuitions about how certain objectors did well in one sense but not another should, if the account is satisfactory, be fully captured. Here are how the results pan out:

	UNSUCCESSFUL SAMARITAN	ATLANTIS ATTESTATION	OPPORTUNE OBJECTION	PUB	PUB*
SUCCESS NORM	No	Yes	Yes	No	Yes
COMPETENCE NORM	Seat: Yes Shape: Yes Situation: No	Seat: Yes Shape: No Situation: Yes	Seat: Yes Shape: Yes Situation: Yes	Seat: Yes Shape: Yes Situation: Yes	Seat: Yes Shape: Yes Situation: Yes
APTNESS NORM	No	No	No	No	Yes

Table 2.1: *Results under new evaluative norms*

Each case is accurately diagnosed. Let us begin with UNSUCCESSFUL SAMARITAN. It was undeniable that Andrew was unsuccessful insofar as he certainly did not change the mind of his denier counterpart, thus he fails the SUCCESS NORM. Nevertheless, we felt that there was a lot he did (epistemically) right—he presented good evidence and solid argumentation, etc.—which the SIMPLE NORM could not explain. The COMPETENCE NORM, on the other hand, accounts for every element. Andrew clearly displays an *inner competence* (SEAT and SHAPE) in his objecting, a positive evaluation. He only lacks a *complete competence* because he was in an extraordinary situation (a Holocaust denial convention) that was antithetical to belief revision. But, as has been mentioned many times throughout, we do not judge someone's competence on highly hostile situations. Andrew crashed his car on an oil spill here, so despite his ultimate

lack of success (and aptness), he is evaluated as doing rather well in the competency of his objection—the exact and correct level of nuance we wanted in the evaluation.

Turning now to ATLANTIS ATTESTATION, Bryony does meet the SUCCESS NORM as Charlie did jettison his climate change scepticism. But it was clear that there were serious problems with Bryony's objection due to inventing this story about Atlantis to make her case compelling. Again, we get a persuasive diagnosis from the COMPETENCE NORM: Bryony evidently knows how to object successfully in general, so she gets a tick for SEAT and has an *innermost competence*. The fact that Charlie changed his mind and the situation was otherwise everyday means appropriate SITUATION. Bryony falls short on the SHAPE front; she did not have the requisite knowledge on climate change to be in proper SHAPE, hence her decision to go down the mendacious path. Without proper SHAPE, however, she cannot manifest a *complete objecting competence* nor meet the APTNESS NORM, thus, despite her success, she is negatively evaluated for lacking an *inner competence*. We again achieve our desired level of subtlety in the evaluation.

OPPORTUNE OBJECTION likewise gets a satisfying diagnosis. There was both success (Ewen no longer believes vaccines kill) and a good objection from Diana (she offered up myriad good evidence in her objection). In the current nomenclature, the SUCCESS NORM and COMPETENCE NORM are met. Do we have our first *apt* objection then? We should hope not, as despite these facts of success and competence, there was an unmistakable sense that *something* was awry in the case. There seemed to be some problematic luck in play that impinged upon the overall goodness of the objection, after all Ewen only ultimately changed his mind due to some rather extraordinary serendipity. This is where the APTNESS NORM comes into play. Of course, this example was an objecting Gettier case (Gettier 1963) with its stroke of bad luck (Ewen's hair-trigger rejection) followed by a stroke of good luck (the petal-evoked memory) which corrupts the connection between the objection and the belief revision. Thus, the APTNESS NORM is not met despite its two siblings being achieved because the objection was not successful *through* the complete competence—luck played a sizeable role. So, we get the fitting analysis that the objection does very well in meeting both SUCCESS NORM and COMPETENCE NORM but the intuition of slight degradation is met by failing the APTNESS NORM.

Finally, the two PUB variations. The first was from the opening of §2.2 where Bella objects perfectly well but Alex is rather confident and remains steadfast in his incorrect belief. I do not think this is an extraordinary SITUATION (a la UNSUCCESSFUL SAMARITAN) antithetical to belief revision, rather this is the fairly common scenario where someone is confidently wrong. Thus, Bella achieves a *complete objecting competence* as she should for her good objection. This may give us some pause. Is it not strange to fully meet the COMPETENCE NORM and still fail? Surely failure always stems from some inadequacy in SEAT, SHAPE, or SITUATION? This is mistaken,

however. In a slogan, reliability does not entail infallibility. Ronnie O'Sullivan has the greatest ever snooker SEAT and could be in proper SHAPE and appropriate SITUATION and still *sometimes* miss a pot. Indeed, it would be a tremendous flaw in the account if meeting COMPETENCE NORM *did not* allow for failure. Thus, this PUB case is a perfectly normal result. Finally, our trivial and simple PUB* where Bella has a nice objection and Alex takes her up on it immediately is the first to meet the APTNESS NORM, getting full marks, as it well should.

The first novel contribution of this chapter is now complete: Sosa's framework of performance-normativity has been applied to objecting and shown to be an elegant and extensionally adequate solution to answering questions of objection evaluative normativity. On its own, this is a respectable achievement. Nevertheless, we will not stop here. I will now show that on top of all the above, the framework can also tell us about the evaluative normativity of a particularly nascent area: *forbearance from objecting*—what I call *quiescing*.

2.7 QUIESCING

2.7.1 From silence to quiescing

We object to others' asserted falsehoods all the time. Nevertheless, *quiescing* in the face of an asserted falsehood—that is, being silent and not objecting—is also something we do. For whatever reasons we may have, sometimes, we are *silent*. This phenomenon has received some philosophical treatment and while I do not think much of it will prove ultimately relevant for this chapter (the reasons for which I will soon get into), it is worth giving a quick overview of the story so far.[20]

Sanford Goldberg (2020, chap. 8–9) has some of the seminal work on silence, which he takes to be, "the state of remaining quiet in the face of a publicly made claim (statement or assertion), wherein the audience gives no explicit indication where he has accepted or rejected the claim." (2020, 153) From this, he proposes his *No Silent Rejection* thesis. Roughly, this is the idea that if someone is silent in response to an assertion, we have a defeasible entitlement to assume that they do not reject what has been said.[21] This also implies that:

> [A]ny audience who is known to have observed an assertion and who rejects, or who otherwise has doubts regarding, the observed assertion has a generic pro-tanto conversation-generated practical reason to give some public indication of this. (153)

At first glance, this and *No Silent Rejection* do appear quite relevant to my project here. In fact, one might think if Goldberg is right, some of the *oomph* has been sapped from

20 Later in this section when I home in on how precisely the evaluative normativity of quiescing works, I will flag where parts of my discussion may prove relevant or of interest to this extant work.

21 For arguments against this idea, see Lackey (2018), Tanesini (2018), and Klieber (2024).

the account given above; the thought being that if we do have this obligation to object, then epistemically good objections end up being in some sense "supererogatory" insofar as the top priority is flagging your disagreement, anything beyond that (i.e., meeting my SUCCESS, COMPETENCE, and APTNESS NORMS) is basically a somewhat unimportant bonus. Moreover, any account of the evaluative normativity of *not objecting* is going to be essentially redundant because, unless the obligation is defeated, you ought not to quiesce in the face of that which you take to be wrong. Nevertheless, I think we should resist these conclusions for a variety of reasons.

First, we should be clear that my objecting and quiescing are not identical with Goldberg's flagging disagreement and being silent in response to an assertion, respectively. Presumably, the disagreement flagging has its own internal aim of, say, *displaying to interlocutors that one thinks the in-play assertion is false*, which certainly pulls it apart from my DOXASTIC ACCOUNT objecting. Being silent and quiescing are similarly different—the details of which are soon coming—insofar as being silent is broader and not necessarily in relation to (what one takes to be) false assertions and one's decision not to object like I take quiescing to be.

Nevertheless, my opponent could press further that the *spirit* of Goldberg's point (being silent has pernicious consequences) is still worrisome for this chapter, particularly when it comes to my extension of the view to quiescing. By objecting there is this potential of upside where your interlocutor jettisons their false belief that is entirely unavailable if one is quiet. So, the best-case scenario for being quiet is as beneficial as the worst-case scenario for objecting—your interlocutor does not change their mind.[22] However, this latter thought is mistaken, and in showing so, the value of the evaluative normativity of quiescing will also become clear.

Consider a final vignette with an objecting variant, DIALECTICAL DEFEAT, and a quiescing variant, COGNISANT QUIESCE:

> Fred is at the wine reception of his sibling's wedding, conversing with new his sister-in-law, Gabrielle. A few glasses in, Gabrielle puts her forward her view that anthropogenic climate change is completely incorrect and indeed invented by the Powers-That-Be for various nefarious purposes.

DIALECTICAL DEFEAT

22 This is arguably partly in the background of *No Silent Rejection* (although Goldberg's main concern is flagging disagreement to third parties). Nevertheless, one might think this is independently plausible insofar as we might have practical and/or moral reasons to not object, but epistemic ones are not going to be decisive. Lackey arguably endorses something along these lines as the negative consequences stemming from objecting she considers are primarily ethical or prudential (2020b, sec. 3). Indeed, in her tetralogy of papers (2018; 2020a; 2020b; 2021) which discuss objecting to various extents, only the penultimate paragraph of the final paper considers potential negative *epistemic* effects (2021, 294–295). The adjacent and often overlapping literature of *engaging* also often tacitly presupposes this general idea; e.g., see Mill (1859) and Cassam (2019b).

> Fred knows this is false and, despite recognising that he does not know much about the ins and outs of climate change and that Gabrielle is extremely confident and stubborn, objects. Unfortunately, Gabrielle appears to have a wealth of (descriptive) knowledge on the topic and is ready for a fight. She attacks him very effectively, tearing apart his evidence and views, and seems to essentially prove him wrong to onlookers.
>
> COGNISANT QUIESCE
> Fred* knows this is false and so considers objecting. However, he recognises that he both does not know much about the science at play, and Gabrielle seems like she is ready to fight her case. In light of these factors, Fred* decides it is best to quiesce instead of objecting.

The general thought here should not be too contentious: the outcomes in the latter variation do not seem as pernicious as the former. DIALECTICAL DEFEAT plausibly carries with it myriad negative epistemic effects: Fred might lose his knowledge,[23] onlookers could similarly have their knowledge defeated, and so on. An exact diagnosis is not crucial; suffice to say, this did not go well and Fred manifestly fails on SUCCESS, COMPETENCE, and APTNESS grounds. In light of this then, it would be odd to criticise Fred's COGNISANT QUIESCE counterpart. In fact, there seems something laudable about this quiesce. Here, Fred* recognised his requisite lack of competence (in particular, his improper SHAPE and the likely inappropriate SITUATION), and in choosing not to object, displayed some admirable reticence. This is not a mystery, however. In fact, as I will now detail, the performance-normative framework that we constructed for objecting can also diagnose these cases and explain the epistemic credit due in COGNISANT QUIESCE.

2.7.2 Proper aims, forbearance, and suspending

We start by looking at the *proper aim of attempts* in general: "to make an attempt on that target if and only if the attempt would succeed aptly." (Sosa 2021, 66) This seems right; a footballer who shot every time they received the ball would be criticised for this performance, even if they occasionally scored. So, to meet the proper aim biconditional, it is built in that one must not make an attempt if it would not succeed aptly. This in turn generates two avenues of failure:

(a) One might make an attempt on the target when one *would not* succeed aptly.

(b) One might fail to make an attempt (on the target) when one *would* succeed aptly. (66)

23 Fantl (2018) discusses this possibility when cautioning against arguing about positions you know to be false.

This is how Sosa manages to bring *suspending* (viz., not settling on the question of whether p) under his telic virtue epistemology—a prima facie difficult task considering his picture is one of *attempts* and *suspending* looks like a paradigmatic example of *not making an attempt*. Nevertheless, (a) makes clear how suspending could be (epistemically) appropriate; if, say, one's evidence equally supports both p and $\neg p$, not only is the trivially correct epistemic action to come down on neither p nor $\neg p$, (a) tells the believer not to make an attempt (i.e., suspend) because even if they believe p and it turns out p is true, this would not be an *apt* belief (knowledge). It would be a lucky guess akin to shooting from the halfway line and scoring.

This what Sosa calls *narrow scope forbearing*: "(Forbearing from *X*'ing) in the endeavour to attain a given aim A." (Sosa 2021, 49) In gnoseology, our given aim is to believe aptly, thus when one suspends because they lack conclusive evidence on whether p, they are suspending *to achieve this proper aim*. This is to be distinguished from *broad scope forbearing*: "Forbearing from (*X*'ing in the endeavour to attain aim A)." (49) This is the criticisable form of forbearance; it is not done in the name of the proper aim, it is forbearance from the entire domain. For instance, one's suspending on p irrespective of what their evidence indicates. And, again, even if it just so happens that suspending *would be* epistemically appropriate, this would not be an epistemically creditable suspension as its ultimate appropriateness was only luckily true and not the catalyst for the suspension. This is an instance of the avenue of failure (b).

Translating the proper aim of attempts into gnoseological terms, we generate two symbiotic aims:

> OVERARCHING AIM
> Answering one's question aptly, with an apt alethic affirmation.
>
> SUBORDINATE AIM
> Affirming alethically iff that affirming would be apt (otherwise suspend).

So, it is now clear how Sosa managed to insert, in a principled way, suspending into his telic virtue epistemology. What this demonstrates is that the performance-normative framework has the acumen to explain forbearances from the performance at hand, thus we can now turn to objecting and quiescing.

2.7.3 The aims of stlīsology

Taking the gnoseological aims from above, we can translate them into the aims for the subject of this chapter, what I am calling *stlīsology*:[24]

> OVERARCHING AIM
> One's interlocutor discarding their false belief from an apt objection.

24 From the Latin *stlīs* meaning "dispute". I am not necessarily married to this neologism but it is a suitable placeholder for now.

SUBORDINATE AIM
Object iff that objecting would be apt (otherwise quiesce).

As I have argued above, the gold standard of objecting is doing so aptly, so it stands to reason that the paramount aim of stlīsology is to attain apt objections, just as the aim of any performance is to do it properly. The biconditional in the SUBORDINATE AIM falls out of this. Thus, we can regenerate the specific dimensions of falling short:

(c) One might object when one *would not* succeed aptly.

(d) One might quiesce when one *would* succeed aptly.

We now have the requisite machinery to explain the cases in which Fred featured and give a satisfactory account of quiescing. Recall in DIALECTICAL DEFEAT that Fred was cognisant of his poor SHAPE (he does not know much about climate change) and the dicey SITUATION (Gabrielle's readiness to litigate and stubbornness). What this means is that even if by some miracle Fred did meet the SUCCESS NORM (perhaps going down the Bryony route from ATLANTIS ATTESTATION), his objection would never meet the APTNESS NORM, and thus would still be bad in myriad ways. Of course, Fred was wholly unsuccessful as well, failing to meet either of the aims, and thus fell foul of (c) by objecting when he would not succeed aptly.

Nevertheless, the evaluative normativity of DIALECTICAL DEFEAT was presumably already handled by our earlier discussion of *objecting*, what is really at stake here is COGNISANT QUIESCE, and Fred*'s quiesce therein. What we want is for Fred*'s quiesce to be epistemically creditable in some way, to be better than his counterpart's inapt objection. This diagnosis is achieved. By accurately tracking his bad SHAPE and poor SITUATION, Fred* is aware that he is not failing on the grounds of (d)—his objection would not be apt *even if* it somehow succeeded—thus his quiesce is epistemically creditable to him as he is following the SUBORDINATE AIM even though the DOMINANT AIM is currently out of his reach.[25] Schematically, then, one's proper quiescing tracks the COMPETENCE NORM in the following way:

> One properly quiesces in the face of a false assertion (aiming to object to it aptly) iff one quiesces based sufficiently on one's lack of the *complete objecting competence* (SEAT, SHAPE, and SITUATION) required in order to object to the false assertion aptly.[26]

Still my opponent might continue to press: why think that this quiescing is "better" than the inapt objecting? After all, both Freds are ultimately unsuccessful—one might think we may as well be hanged for a sheep as a lamb, as it were. Moreover, if we buy

25 Goldberg's *No Silent Rejection* thesis is defeated if one is not in a "cooperative exchange." (2020, 175) Plausibly, the picture I am constructing here aligns with this idea insofar as an uncooperative exchange could be a case of inappropriate SITUATION for a *complete objecting competence*.

26 Cf. Sosa (2021, 94–95).

this picture of proper quiescing, it entails that both Andrew from UNSUCCESSFUL SAMARITAN and Bryony from ATLANTIS ATTESTATION would have had epistemically laudable quiesces despite the fact that we positively evaluated Andrew to an extent, and Bryony was actually successful.

We can make good sense of all these diagnoses, however. First, consider Sosa on why proper suspending is better than inapt believing:

> [S]uspending now may be a means to an apt answer to that question in due course, and a thinker may suspend as a spandrel outcome of properly pursuing that objective, and also as a means to eventually attaining it. (Sosa 2021, 93)

The idea is not merely that inapt believing (or indeed an inapt objection) is in direct contravention of the DOMINANT AIM, it is that *by* making such inapt attempts, it will take one *further away* from the DOMINANT AIM in general. For Bryony, it is easy to see that her mendacious objection is problematic, irrespective of her ultimate success, and this ought not be epistemically encouraged—in the same vein that we do not find it surprising to criticise the belief and method where one simply guessed and was correct. But, on top of this, Bryony was essentially extremely lucky that her objection went through and was successful insofar as randomly making things up is not a reliable means to object well, and is in fact more likely to be a reliable means to having people *remain steadfast* in their false beliefs (i.e., suppose Charlie was aware that Bryony was lying about Atlantis, it would plausibly make him *more* secure in belief about the unreality of climate change as opposed to less). Instead, it would have been creditable to Bryony had she quiesced and, following a spandrel outcome, perhaps improved her SHAPE *and then* objected with a *complete objecting competence*. Similarly, Andrew could have properly quiesced, and waited for a better SITUATION as a means to eventually attaining (or at least trying for) an apt objection.

This is even clearer with DIALECTICAL DEFEAT. By objecting so inaptly, Fred has not just missed both the DOMINANT and SUBORDINATE AIMS, he has made it such that Gabrielle is likely more secure in her false belief due to her dialogical slaughter of Fred, and now her SITUATION is potentially even more antithetical to belief revision. The DOMINANT AIM is more out of reach than ever. Fred* in COGNISANT QUIESCE has none of this; he now has the possibility to improve his SHAPE, wait for an improved SITUATION; but, crucially, he has not done anything to take the DOMINANT AIM further away from being achieved. Fred* has in fact properly pursued his objective of objecting aptly and thus his quiesce is apt.[27] Thus, we can see why proper quiescing is preferable to inapt objecting: it keeps the DOMINANT AIM in sight and

27 To be clear, none of this is prescriptive in the sense of being categorical norms about what Fred (or any of our other protagonists) "(epistemically) ought to do" as Lackey/Johnson were interested in giving. These are simply *evaluations* that relate to the aimed action that the objectors have undertaken.

positions one to eventually attain it in a way that is not possible if one opts to go for an objection that lacks a *complete competence*.

In sum, then, not only do we have an extensionally adequate and elegant framework for evaluating *objections*, we can use that same framework to evaluate objecting forbearances. This latter part is especially novel: an epistemically creditable quiesce is new territory for this literature on silence.

2.8 CONCLUDING REMARKS

Sosa distinguishes between *gnoseology* and *intellectual ethics*: the former is the familiar theory of knowledge and evaluations of belief as apt or not therein; the latter is more concerned with the ethical or practical dimension of such knowledge and beliefs. The classic illuminating example is counting blades of grass in one's lawn. One could do this successfully through their complete competence, and so aptly, and thus be knowledgable, receiving full marks from the performance-normative framework. Nevertheless, this knower can still be critiqued for pursuing pointless ends and wasting their epistemic labour—this critique stems from the intellectual ethics side.

This chapter's focus has been on the epistemic side of objecting and quiescing, what I (tentatively) labelled stlīsology. Stlīsology's intellectual ethics analogue we could call the *ethics of engagement*. And indeed there will no doubt be many blades-of-grass-case analogues. Every time someone tells you the time and rounds up or down slightly, you could object aptly and hit the gold standard of stlīsology. Nevertheless, that is where the ethics of engagement kick in; one would rightly be criticised for such pointless and probably aggravating objections—in spite of any stlīsological kudos garnered. Reversed cases will no doubt be common as well; instances where one speaks up for moral reasons and does not ultimately do so aptly. Had Andrew quiesced we could have given him stlīsological credit, but perhaps he would have been *morally* blameworthy. None of these are bullets to bite, they are simply different domains of evaluation. My focusing here on stlīsology is not some implicit value judgement either insofar as I am not claiming that the epistemology of objecting and quiescing is more important or supersedes the ethical or practical dimensions. Rather, the epistemic side of this has been sorely neglected by the extant literature, particularly how we would epistemically evaluate such quiesces and objections, hence my aim of developing this comprehensive picture of the evaluative normativity of stlīsology. A task which has now been completed in a way that, I hope, was successful, competent, and apt.

Linking Interlude II

In the previous chapter, I argued that an epistemically good objection is an apt one, building a complete picture of the evaluative normativity of objecting based on a Sosan performance normative framework. I then extended the view to consider *forbearance from objecting*, what I called quiescing. Near the end of the chapter, I considered an instance of quiescing that I felt entitled the quiescer to some epistemic credit for their prudent decision. Part of the justification for I mentioned for the laudability of this quiesce was, if the quiescer had instead gone ahead and objected and it had gone terribly wrong, it could have bad epistemic consequences, not only on the objector themselves, but on potential *bystanders*. This following chapter takes this line of thinking up in full, broadening the focus from merely the engager and the engaged as the previous chapter did. In particular, I will show that, when it comes to a certain type of claims I call *controversial false assertions*, it is incredibly risky to engage and object to such claims due how often it can be wholly unsuccessful and the widespread deleterious epistemic effects on third-parties when it is.

Chapter 3

On the perils of engaging

Recent work in social epistemology has discussed obligations to engage with challenges to our beliefs like climate change denial or anti-vaccine sentiment, and the potential benefits to and dangers for both the engager and the engaged from doing so. The spotlight being trained here, however, has elided a key issue: the possible risks from engaging for third-party observers, not merely the engager and the engaged. In this chapter, I argue that not only are these risks an underappreciated aspect of engaging that ought to be discussed, their neglect is particularly notable because the potential negative epistemic fallout threatens to overwhelm any possible benefits that may be gained from engaging, regardless of how the engagement actually goes. I close by drawing out the theoretical and practical implications from this and sketch a few strategies to conceivably avoid said risks.

3.1 Introduction

What can one do when an interlocutor asserts that climate change is a hoax? What *ought* one do when an interlocutor denies the Holocaust? Why should one engage with an anti-vaxxer? Why should one *not* engage with a flat Earther?[1] These are pertinent and difficult questions, and ones that certain philosophers have recently taken up in the social epistemology literature (McCormick 2023; McKenna 2023; Cassam 2019b; Fantl 2018; Lackey 2020b; Battaly 2018a, 2018b, 2021). Naturally, there is disagreement between those working on these areas, but it is illuminating to highlight where the spotlight has been focused and the trends of agreement that can be flagged.

First, a key reason for engaging with those who believe and assert falsities is the hope that by doing so, one might be able to change their mind, and get them to update from a false belief to a true one. As Quassim Cassam writes: "If one can't be both-

1 To be clear, in this chapter, I am focused on citizen/peer interactions in day-to-day circumstances and not the *scientific experts/lay people* interactions that characterise, for instance, Nancy Rosenblum (2020)'s "witnessing professionals" and their plausible duties to speak out and engage with climate change deniers.

ered to argue against conspiracy theories one can hardly complain if people end up believing them." (2019b, 117) This thought is a driving force behind the literature (in particular those noted above) and understandably so; after all, however sceptical one might be of the prospects of such a result, it is plausible that there is at least a *chance* of changing someone's mind for the better, and this chance is presumably much higher conditioned on engaging than not.

It is perhaps unsurprising, then, that some view engaging with challenges to our beliefs as *obligatory*. John Stuart Mill's *On Liberty* (1859) is really the *locus classicus* of this thought: he argues that only by engaging with those we disagree with can we garner a myriad of positive effects—ensure that what we think to be true really is so, prevent commonly held beliefs from becoming unimpeachable "dead dogmas", and gain further understanding of our own reasons for our beliefs.[2] He even thinks engaging with opinions that are false can be useful as they very often "contain a portion of the truth," (118) which can only be appreciated and taken up via such engagement or discussion.[3] Cassam (2019b) argues along similar lines, saying that if we do not engage with challenges to our beliefs, we lose our right to our justification of them, and thus lose our knowledge, hence the obligation. The idea is that if one cannot respond to a counterargument to one's belief, then one is simply not justified in holding it, and so it no longer constitutes knowledge.[4] Jennifer Lackey (2020b)'s duty to object is in a similar ballpark inasmuch as she takes it as an important factor to flag our disagreement to help prevent bystanders from taking up the asserted falsity, although she downgrades the obligation somewhat to an *imperfect* duty where one must only engage sometimes, and how great this normative pressure is depends on a couple of different social and practical elements. McIntyre (2018, 2021) also acknowledges this point about public dismissal of claims helping to insulate against those coming across such ideas for the first time getting the wrong impression—i.e., that they are orthodox opinions.

The foregoing discussion essentially captures the key (epistemic) positives of engaging when it goes well. We can break the potential avenues of positive effects down into three broad areas: (i) positive effects for the engager (the one replying to the individual making the problematic claim, (ii) positive effects for the engaged (the one initiating the conversation by making the problematic claim), and (iii) positive effects for third parties or society as a whole. For (i), we have a gain in the understanding of one's own reasons for their beliefs and learning the whole truth about the topic at hand (Mill 1859), the right to maintain our knowledge (Cassam 2019b), and/or

2 See, for instance, McKenna (2023, chap. 6), Fantl (2018, chap. 5), or Macleod (2021) for more on Mill's defence of an obligation to engage.

3 While this may well be true in some (or even most) ordinary circumstances, I do not think it is particularly applicable to the cases (soon-to-be) employed in this paper. Indeed, Mill's thought here implies that there may be truth on both sides of, say, the Holocaust and its denial, which strikes me as rather dangerous. My thanks to an anonymous referee for urging me to discuss this.

4 See Aikin (2010) for more on such a dialectical requirement on justification.

the fufilment of an epistemic duty (Lackey 2020b). For (ii), we have the potential of their updating to a true belief (McCormick 2023). And for (iii), we have a general increase in the stock of true and secure beliefs in the epistemic environment (Mill 1859; McKenna 2023) and an attempt to impede bystanders from taking up the false belief (Lackey 2020b; McIntyre 2018, 2021).

Not everyone is so optimistic about the overall prospects of engaging with challenges to our beliefs, however, and have turned to discussing the possible *negative* epistemic effects from doing so.[5] Jeremy Fantl (2018), for instance, is sceptical of this strong Millian position of open-minded engagement. He argues that such engagement with that which we know to be false can in fact lead to *losses* of knowledge for the engager as their confidence may be distorted if they cannot respond effectively to their interlocutor. The thought is that even if one knows an argument against, say, the reality of the Holocaust is misleading, an inability to effectively respond to it and demonstrate precisely where and why it goes wrong can adversely affect one's confidence in their knowledge of the Holocaust's reality, and potentially even result in a loss of said knowledge. For such reasons, then, he argues that one ought not engage open-mindedly with asserted falsehoods to try to insulate against this concern. He does think that we can perhaps engage *closed-mindedly* with false assertions—that is, engage without the possibility of changing our mind—but is even somewhat sceptical of the prospects of this strategy. McCormick (2023), Battaly (2021), and McIntyre (2020) agree with Fantl about those dangers of open-minded engagement but are more optimistic at the prospects of the closed-minded variety, advocating for this strategy as a key method for changing the minds of those who believe things like climate change hoaxes or Holocaust denial. Once one is closed-minded to the possibility of *their* mind changing in response to discussion with a fringe believer, the thought goes, all possible risk from engaging has been eliminated (in terms of the engager's own knowledge at least), and the only way is up, inasmuch as there is now a chance of changing the fringe believer's mind to a true belief—or at least the possibility of their jettisoning a false one.

McKenna (2023) has a different worry with Millian and Cassamian open-minded engagement; he argues that some stand to lose a lot more than others by engaging open-mindedly with those that disagree with them—specifically those who may suffer a *testimonial injustice* if they engage.[6] The idea is that, say, a woman of colour may

5 For a different discussion of the potential costs from engaging in general (and indeed whether to even engage or not in the first place), see Paglieri (2013). While he does discuss "negative consequences", he is not narrowly interested in epistemic ones as I am, and is generally more focused on *practical* costs such as putting strain on one's relationship with the interlocutor at hand (156) or "social image [and] self-esteem." (160) Nevertheless, it is interesting that we both employ a sort of consequentialist, expected value framework for looking at arguments and engagements with opinions contrary to our own.

6 See, of course, Fricker (2007) for the classic work on epistemic injustices. See also Terzian and Corbalán (2021) for a discussion of potential testimonial injustices suffered specifically with relation to vaccine hesitancy/safety denial.

be shut down if she testifies to her male boss about, for instance, her experience of sexual harassment in the workplace, and thus may end up losing her knowledge of what happened due to her not being respected as a knower. Thus, McKenna argues, she *cannot* have an obligation to engage with such a challenge to her beliefs.

Possible negative effects on the *engaged* from entering into a discussion have not received much of a philosophical treatment—which is not overly surprising. After all, the literature is generally focused on engaging with those who express false beliefs, thus they are already in a bad situation. How things might get *worse* for them clearly has not been much of a priority (in the philosophical discourse, at least). Nevertheless, there is work in the social psychology and political science literatures that bears mentioning: backfire effects where being presented with countervailing evidence can result in one actually being *more* secure in their (false) belief (D. H. Cohen 2005; Nyhan and Reifler 2010, 2015), and polarisation which again can make said beliefs more extreme and secure (Sunstein 2002).

This discussion captures the core strands of the literature so far in terms of the possible *negative* effects stemming from engaging. For the engager, they may lose their knowledge if they engage in an open-minded way and find that they do not have effective responses to the engaged's arguments, or they may suffer epistemic injustices depending on their social situation. For the engaged, they may become further entrenched in their problematic beliefs. This is the dark reflection of the potential *positive* effects I outlined above. Something interesting to note, however, is that there is one area that lacks a negative counterpart: the epistemic effects on other people outwith the engager and the engaged—viz., third parties and/or society as a whole. It might be helpful to present this visually:

	Possible positive epistemic effects	**Possible negative epistemic effects**
Engager	Fulfil an epistemic duty (Lackey), learn the full truth and your own reasons better (Mill), maintain right to knowledge/justification (Cassam)	Lose knowledge (Fantl, McCormick, Battaly, McIntyre), suffer epistemic injustice (McKenna)
Engaged	Take up true belief/get rid of false belief (All)	Polarisation (Sunstein), backfire effect (Nyhan and Reifler)
Third-party observers/ society	Increase stock of true beliefs in society (Mill), prevent bystanders from taking up false belief (Lackey), prevent confusion in others (McIntyre)	?

Table 3.1: *Engaging literature overview*

This chapter aims to fill the question mark cell.[7] I do not want to fill this lacuna for mere posterity, however. I will in fact argue that it is especially concerning that this area has been so neglected in the literature as it has the potential to be by far the most problematic in terms of negative epistemic effects. Indeed, so much so that it threatens to overwhelm any of the possible positives that have been so far discussed. Central to this idea is that much of the literature (in particular, those cited in the *Engaging literature overview* table) has elided key questions relating to *how* engagements can actually go; per the above, one could be forgiven for thinking that engaging is often simply a quick dismissal of "that's wrong" which either works in changing the target's mind or fails, then we all get on with the rest of our day. Engaging is often far more dialogical than that,[8] and this gives scope for it to go far more wrong than has previously been appreciated by many of these authors.

Specifically, I will here focus on engaging with what I am calling *controversial false assertions*. I will not precisely define these but the class of assertions I have in mind are

7 Fantl (2018) discusses an adjacent area in the final chapter of *The Limitations of the Open Mind*. Even so, he is mainly focused on (psychological and intrinsic) harms from inviting what he calls "problematic speakers" to university campuses. This is similar to a worry Levy (2019) raises about platforming problematic speakers creating misleading higher-order evidence about their credibility. These are quite different concerns from the sorts of cases I will discuss and the more (first-order) *epistemic* worries I'm interested in. In short, I am not thinking about platforming specifically, nor psychological or higher-order evidence issues.

8 Again, see Paglieri (2013). Cf. Dutilh Novaes (2023)'s philosophical model of argumentation.

ones such as "climate change is a hoax", "the Holocaust didn't happen", or "Covid vaccines are deadly". These examples all overlap with Neil Levy's *bad beliefs*: "a belief that (a) conflicts with the beliefs held by the relevant epistemic authorities and (b) held despite the widespread public availability either of the evidence that supports more accurate beliefs or of the knowledge that the relevant authorities believe as they do." (2021, xi) Nevertheless, I take my *controversial false assertions* to differ insofar as "bad beliefs" for Levy are not always false—unlike the cases I am focused on. He also generally takes those who believe bad beliefs to be rational which is not something I want to commit to, nor is it important for my purposes here.[9] Plausibly, the closest comparison to my *controversial false assertions* are what Fantl calls "controversial propositions" for which he gives examples of "the theory of evolution, psychic phenomena, the efficacy and danger of vaccines, convoluted conspiracy theories, repugnant moral positions, the existence of God, and whether the Holocaust occurred." (2018, 28) As he is interested in engagers losing *knowledge* in relation to these subjects, he presumably takes the controversial propositions of interest to be false, thus aligning more closely with me than Levy did. On the other hand, his inclusion of "psychic phenomena" and "the existence of God" is certainly outwith the bounds of my discussion here, so I do not want to adopt this conception wholesale. Overall, I think the key point here is that I will follow *both* Levy and Fantl to an extent in not worrying about giving precise conceptual analysis of my umbrella term for the examples employed in this paper. Instead, we can leave the general category at a more intuitive level, as the arguments and examples themselves are of greater importance than any definition, while noting that the exact characterisation likely lies somewhere around those of Levy and Fantl just discussed.[10]

Per the extant literature, one gets the sense that, if the engager is sufficiently closed-minded and any threats of epistemic injustice are accounted for, we have exhausted the ways in which engaging with or objecting to such controversial false assertions can go epistemically wrong.[11] All that is left now is either the engaged changes their mind or they maintain their false belief—and in the latter case nothing has been lost anyway. I will show that there are in fact a variety of ways that such a discussion can go, where, for instance, an engager can 'lose' the engagement, and that this can have wide-ranging negative epistemic effects not even primarily on the engager and the engaged but rather *third-party observers* and *society at large*. Moreover, the epistemic position prior to engaging with a challenge to our beliefs is such that it is very difficult for a potential engager to know how such a discussion might go,

9 Granted this diagnosis does follow if you are working with a subjective Bayesianism framework as Levy is, but I do not want to commit to that either.

10 My thanks to an anonymous reviewer for urging me to give more detail on this.

11 Again, my focus here is purely on the epistemic dimension. There are of course many *non-epistemic* things that can go wrong (e.g., violent reactions and non-epistemic issues of respect that cannot be accounted for under the epistemic injustice framework). Once again, see Paglieri (2013) for more examples of potential non-epistemic costs.

meaning that even the mere potential of a catastrophic engagement threatens to overwhelm any potential benefits that might be garnered in a successful engagement. In this sense, my argument here is certainly a pessimistic one insofar as I will argue that engaging with *controversial false assertions* is an extremely risky activity and one that we should be very reticent about doing in a lot of ordinary circumstances. Nevertheless, all hope is not lost, as I will close with some suggestions of how plausibly avoid these problems.

3.2 CASES

I will now go through a few different cases, all of which have the same structure: someone putting forward a *controversial false assertion*, someone engaging with and objecting to it, and an audience of third-party onlookers. These cases are primarily designed so as to make it easy to see the intuitions being pumped and the arguments being drawn from them, but they have the added virtue of being closely analogous to any such discussions that might take place *online*—i.e., on social media. I take it that this is a not insignificant asset as we might think that a lot of the time people are faced with *controversial false assertions* is in these settings, where there is an such an audience of onlookers, so any conclusions drawn here from the cases should apply more broadly than merely the exact in-person scenario discussed in the vignette. In particular, such conclusions are relevant to online discussions because the general set-up will often be very similar insofar as there is an audience not participating but bearing witness to the engagement (viz., the users scrolling past such a discussion) and then the people actually engaged in the argument. Moreover, as my focus in this chapter is primarily on third-party bystanders, and their role—whether online or in-person—is by stipulation non-participatory, then drawing analogies between the two is both reasonable and legitimate.

Each situation will have a different outcome: I will consider a case where our protagonist (the engager) 'loses' the engagement, one where we might call it a stalemate, and one where the protagonist appears to be successful. In the first two, I will argue that we have good reason to be concerned about the negative epistemic effects of such scenarios, while even the success case is not the unadulterated good we may have initially thought. Finally, let us stipulate that these are what McCormick and McKenna might think of as "dream scenarios" insofar as we will assume in each case that the engager is sufficiently closed-minded so as not to risk their own knowledge,[12] and any threats of testimonial injustice are not relevant. The idea is that even with all

12 One might argue that we cannot simply *choose* to be closed-minded like this, and repetition and fluency effects (see, e.g., Levy (2017)) can result in false belief uptake or loss of knowledge irrespective of the engager's intentions. This may well be true, and if so is merely more grist for my mill of the perils of engaging, but I nevertheless set aside such worries here as it is all the more impressive that my argument goes through *without* needing to wield this extra psychological weaponry.

these positive measures secured, there are still vast concerns about negative epistemic consequences. Now that we have all our pieces in place, let us begin with the opening case.

3.2.1 Losing

> DINNER
> Alex is out for dinner with some colleagues. During the mains, one colleague, Bob, starts espousing the view that climate change is a hoax. He presents some 'evidence' and generally puts forward the case of climate change denial well. Alex knows this view is false and so engages with Bob, objecting to his false assertion. Unfortunately, Bob has come rather well prepared and responds very effectively to Alex. He attacks her arguments cleverly, rejects her evidence, provides 'evidence' that she does not know how to refute, and generally seems to prove her wrong (at least from the perspective of the onlookers).

First, it does not strike me as at all controversial to say that this engagement did not go well—in fact, it is hard to imagine how it could have gone worse. Alex engaged because she knew Bob's assertion was false and perhaps hoped to change his mind, but instead found her own view attacked effectively and forcefully in such a way that she was incapable of addressing. Now, we have already stipulated that Alex's engagement was a closed-minded one, so her belief or knowledge that climate change is real remains unshaken. We also stated that there was no risk of epistemic injustices taking place either. Where, then, might we locate the negative consequences that seem intuitively endemic in this case? One might suppose that Bob will become even more secure in his false belief after such a discussion, but he presumably already believed it strongly enough to assert it to an audience, so this problem does not seem particularly pressing either. Our real concern, I propose, relates to the *third-party observers*.

If we assume that the others at dinner follow general societal trends, then most believe that climate change is real (around 70%), some are unsure or suspended (around 15%), and some are in agreement with Bob (also 15%) (Tranter and Booth 2015; Ballew et al. 2019). While we stipulated that *Alex* was closed-minded when she opted to engage with Bob, there was no such guarantee for the others present. Therefore, when Alex's arguments were effectively attacked, and her evidence rejected, and so on, we can easily predict that this could shake any of the third-party observers' beliefs or knowledge inasmuch as their justification for their belief or knowledge could be defeated by these (misleading) arguments.[13] Additionally, those that were unsure or

13 One might suppose that audience members will all react differently to the exchange, so—without more detail on audience biases, attributions of authority etc.—we have no way of knowing who 'won'. Moreover, someone's perception of who 'won' is different from finding the arguments convincing and/or having one's beliefs shaped by the debate. Starting with the latter point, while I agree

suspended now have good reason to come down on one side of the argument—after all, they just saw the position of climate change denial strongly come out on top in a public engagement.[14] Moreover, this is all arguably *rational* for these parties. In fact, if we want to argue reasons *why* we ought to or can engage with such assertions are to change the interlocutor's mind and/or prevent bystanders from getting the wrong impression (in other words, *appeal to their rational capacities*), then there is no principled reason to suggest the reverse cannot be true if the discussion goes poorly—viz., these third parties losing their knowledge or even coming to believe the *opposite* of what we wanted when we engaged.[15]

What upshots can we take from this? Well, with this case we are already in somewhat uncharted territory insofar as none of the extant scholarship really discussed *how* an engagement could actually go, but it is quite clear that the foregoing literature does not really have the resources to explain this case anyway. McCormick and McKenna's worries have been accounted for with closed-mindedness and no epistemic injustice, respectively. Lackey's duty is a deontic one so presumably she is not overly concerned with actual consequences or outcomes. Mill's diagnosis here would likely be a rather odd one in that I suspect he would have to say that everyone should lower their credence in climate change's reality following such a discussion. Cassam almost seems to reject this sort of scenario outright as he apparently thinks it is rather easy to argue against people like Bob.[16] Fantl comes closest to saying *something* about such a case, but even he is more focused on the effects on the *engager* (Alex,

that perceptions around winning and changing of minds can come apart, this will not be the standard case. This is borne out in empirical data; e.g., Aalberg and Jenssen (2007) which found that "the outcome of the debate in terms of who was perceived as winner made a difference [to beliefs around issues]." (131) Or L. Wang et al. (2017), whose model predicts whom audiences ultimately vote as winning a debate, find that "the inherent persuasive effect of argument content plays a crucial role in affecting the outcome of debates" (220) and that "Our model also shows that winners use stronger arguments." (229) Thus, coming back to the former objection, empirical data suggests we do have ways of knowing who won—and these ways relate specifically to the strength of arguments in play. Further empirical data concurs on this: Ettensperger et al. (2023) found that "predispositions may be a strong filter but are far away from being the only determinative factors for the perception of the audience" (14) while Nwokora and Brown (2017), employing as a case study the Obama/McCain 2008 debate, found that when a debate win is convincing, it breaks partisan lines. Thus, we can know who won a debate (audiences are fairly homogenous on this), and who won will generally be determined by strength of argument, as my vignettes assume. My thanks to an anonymous referee for pressing me on this point.

14 There is an interesting parallel here with a discussion Timothy Williamson has in his *Philosophical Method* (2020): "When two senior philosophers argue some issue out with each other in public, with prestige at stake, it is often clear that neither of them will ever persuade the other; even so, it is not a waste of time if there are uncommitted students in the audience, making up their own minds as to which of the two is having the better of the argument." (41)

15 For simplicity's sake, I am not going to precisely discuss credences here, instead just focusing on belief, justification, and knowledge. Note, however, that a lot of this discussion could equally be couched in terms of bystanders lowering their credence in climate change's reality and raising their credence in climate change being a hoax—clear negative epistemic effects.

16 Recall the quote from §3.1: "If one can't be bothered to argue against conspiracy theories one can hardly complain if people end up believing them." (2019b, 117) See also chapter 5 of *Vices of the Mind*.

here)—and we already stipulated that she will not suffer any negative epistemic consequences here.

All this to say I think it is rather clear that cases like DINNER highlight an important dimension of engaging with *controversial false assertions* that has clearly been missed so far in the literature. Not only is it distinctly probable that third-party observers will suffer the sort of negative epistemic effects outlined above, such effects are also far more concerning than the worries the literature has previously touched on just in terms of pure *volume*. Even if only two people in the DINNER audience lose knowledge or inculcate a false belief, that is already more epistemic damage than if we only talk about potential effects on the engager, so this is clearly a problem that we must discuss and take seriously.

Nevertheless, as an immediate rejoinder, my opponent might suggest that this situation where the engager undergoes such a dialectical trouncing like DINNER is rather rare.[17] Or, at the very least, there are certainly more ways a discussion can go than the way we just looked at. So, let us now consider a different variation of DINNER and see what conclusions can be made.

3.2.2 Stalemate

> DINNER*
> Claire is out for dinner with some colleagues. During the mains, one colleague, Derek, starts espousing the view that climate change is a hoax. He presents some 'evidence' and generally puts forward the case of climate change denial well. Claire knows this view is false and so engages with Derek, objecting to his false assertion. They have a long back and forth, ending at a sort of impasse from the perspective of the third-party observers, where neither presents a slam-dunk argument or unanswerable piece of evidence (and neither concedes nor proclaims the other to be correct).[18]

17 As it happens, I am rather sceptical of this thought. Studies have shown that climate change scepticism among US republicans tends to *increase* with greater scientific literacy (Funk and Kennedy 2020), therefore we can reasonably infer that those asserting such falsehoods about climate change will frequently have a lot of descriptive knowledge about climate science and related matters, and thus may well be in a good position to argue their case effectively.

18 One might think that it is easy from the point-of-view of the universe to call Claire and Derek's discussion a tie (or indeed the others wins or losses), but bystanders could draw their own conclusions about which side came across better and so think Claire 'won' and plausibly infer that climate change is clearly real. I think this is right, but obviously the reverse applies too: a bystander could equally think that Derek 'won' and draw the conclusion that climate change is indeed a hoax. Thus, let us assume these two situations cancel each other out. More generally, I will assume that bystanders' idiosyncrasies are not frequently biased in either direction and so will set aside this worry of audiences not assessing evidence homogenously. On the other hand, however, this may be too charitable to my opponent insofar as remaining 'undecided' on a controversial issue (i.e., one on which there exists a broad societal consensus) suggests that the bystander already holds strong beliefs in favour of the false assertion. Thus, there is even more reason to think that a Stalemate will favour the viewpoint of the *controversial false assertion*, and so the perils of engaging are even greater. Either way, my

Starting with the obvious, this outcome is clearly not as bad as the previous case. Nevertheless, I argue that it is still not good enough and remains well placed to have negative epistemic effects on the third-party observers. To see why this is so, we have to consider the point of our engaging with false assertors. From the discussion so far, we can draw out three discrete but related elements: (i) to (hopefully) change the interlocutor's mind, (ii) to display our own disagreement,[19] and (iii) to prevent any third-party observers from getting the wrong idea about your own beliefs and/or hopefully inhibit their taking up of the false belief.[20] These are all connected by the central thread of doing everything we can to limit or prevent inculcation of false belief. The problem is that when the truth (in this case, climate change's reality) goes head-to-head with a fringe falsity (climate change denial), it does not suffice to come out on level pegging. Just as if Partick Thistle, a small Scottish football team not even in the top flight of the Scottish leagues, were to hold Real Madrid, historically one of the greatest teams ever and still one of the best in the world, to a 1-1 tie, fans of Partick Thistle would be well entitled to view this as essentially a win, so too can a deadlocked discussion come out on the side of the climate change denier.

To continue this analogy somewhat, suppose that each of Claire and Derek's 'goals' were an argument that they both responded to very effectively—so much so that any bystander whose belief was justified by this argument had such justification defeated. The problem here oddly comes from the fact that most people *do* believe in climate change; if we equally apportion defeat from the two 'goals' to the bystanders, there are simply more people to lose justification (and thus knowledge) of climate change's *reality* than its denial. For instance, say there are twenty bystanders. Using our percentages from the previous subsection, that is fourteen believers, three suspended, and three non-believers. If we say half of each side have their justification defeated, then that is seven people on the side of climate change who suffer loss of knowledge compared with only one or two who jettison their false belief. Therefore, this dialectical tie clearly favours Derek's side. And, once again, this is all presumably quite rational for everyone—there is no principled reason to think that belief in climate change is somehow more secure or less vulnerable to defeat than the opposing position, I am looking at this in the most fair way possible. Perhaps we could say that *neither* side will shift their beliefs at all, but if we go that far then we have to wonder what the point is of ever engaging with anyone anyway.

So, this stalemate scenario is likewise a serious concern when we engage with *controversial false assertions*, carrying the potential for widespread negative epistemic

argument is supported. My thanks to two anonymous reviewers for urging discussion on this.

19 See also the literature on the psychological function of disagreement in terms of signalling group membership or protecting ego/identity—e.g., Tappin, Pennycook, and Rand (2021).

20 This is reminiscent of S. C. Goldberg (2020) discussed in the previous chapter in §2.7. Nevertheless, his *No Silent Rejection* is a practical reason generated norm and is mainly about *flagging disagreement*. This is quite different from the epistemic norms and consequences stemming from more dialogical discussions that I am interested in here.

effects which again vastly outstrip the epistemic benefits. Moreover, we do not even have the possible comfort (as we did in the previous scenario) that this sort of situation is rather rare; in fact, a back-and-forth discussion that peters out at an impasse strikes me as a rather common outcome. For instance, a recent study found that, "the vast majority of [arguments on Facebook] (roughly 71%) ended in a neutral manner, without reaching any agreement." (Cionea, Piercy, and Carpenter 2017, 444) Combined with general worries about arguments not changing minds (Gordon-Smith 2019) and polarisation (Sunstein 2002), it is reasonable to assume an argument without any real resolution like DINNER* is going to be fairly frequent.

Let us now turn to the final variation of our DINNER cases, the "success story", and see what can be drawn from it.

3.2.3 Winning?

> DINNER**
> Erica is out for dinner with some colleagues. During the mains, one colleague, Frank, starts espousing the view that climate change is a hoax. He presents some 'evidence' and generally puts forward the case of climate change denial well. Erica knows this view is false and so engages with Frank, objecting to his false assertion. Fortunately, Erica has come rather well prepared and responds very effectively to Frank. She attacks his arguments cleverly, rejects his evidence, provides evidence that he does not know how to refute, and generally seems to prove him wrong (at least from the perspective of the onlookers).

At last, we have arrived at the good scenario. Erica has 'won', the case for climate change has been forcefully shown, and so we might reasonably assume that positive epistemic effects will abound. At the very least, it seems that there are no obvious negative epistemic effects here like there were in the previous cases: none of the bystanders lose justification or knowledge, and some or even all of the non-believers jettison their false belief and take up the true one of climate change's reality (although let us assume, like Alex in the first case, the foil here in Frank does not change his mind either).[21] Despite all this, I will now put forward some suggestive comments about a lingering possibility of *some* problematic epistemic effects.

Climate change denial is a fringe position, so the mere fact that it was seriously

21 My opponent might be concerned about the idealisations/simplifications in play here in the vignettes and lack of discussion of background beliefs of the audience a la Beaver and Stanley (2023) who look at the difficulty of getting people to accept propositions that clash with their "core ideology." (156) I want to be clear that, following the extant literature (i.e., those cited in the opening paragraph of this chapter), my focus is on similar ordinary circumstances and general members of the laity who can and will change their minds in response to arguments, engagements, and objections, not, e.g., those for whom climate change denial (or its reality) is a part of their "core ideology" and thus would be unmoved irrespective of how the engagement goes.

engaged with looks to legitimate the discussion in that there was *something* to talk about. If I were to see two people with an interest in physics having a heated discussion about whether light can propagate through a vacuum with or without the presence of a luminiferous aether, I could well assume that this is an ongoing debate for which there is no settled answer.[22] Even if one side looks to come out on top and prove the other wrong, it does not necessitate that they are right, nor that this is a settled issue (as we saw in the original DINNER case). Were people in their day-to-day lives to hear *genuine* debate on climate change denial, see discourse on Covid vaccines being deadly, witness dialogue on the Holocaust's truth—*even if* these sides are being proven wrong—the mere fact that there exists discussion goes a long way in normalising and legitimising the position.[23]

In a sense, the climate change denier has nothing to lose; these sorts of *controversial false assertions* thrive on engagement, and this is a known phenomenon.[24] We can draw some parallels here with actual strategies employed by tobacco (Oreskes and Conway 2011) and oil (McMullen 2022) companies in the past. Promoting a false case is not always about getting people to believe the *opposite of fact* (viz., smoking *does not* cause cancer, or climate change *is not* happening), rather it is often simply about sowing seeds of doubt, causing people to wonder and entertain possibilities that they otherwise would not have considered.[25] Of course, I already stipulated that we will not consider any losses of knowledge or justificatory defeat here with regards to DINNER**, but merely getting people talking about an issue that *is* a genuine matter of fact as though it might not be settled is, I propose, still a significant concern—which is presumably precisely *why* it was a strategy employed by the aforementioned tobacco and oil companies. Now, to be clear, I am not implying that when the current antagonist, Frank, (or, indeed, Bob or Derek from the earlier cases) espoused his false belief, it was part of some grand strategy in the way tobacco companies tried to fudge the link between smoking and health issues, but I contend that the general outcome *is* similar

22 I would be wrong about this; the luminiferous aether hypothesis was conclusively discarded in 1887 due to the Michelson-Morley experiment. See Shankland (1964) for the full account of events.

23 Theel, Greenberg, and Robbins (2013) discusses how CBS overrepresented climate change deniers in the news by around six times their actual representation in the scientific community. It is plausible that there is a causal connection between this and the fact that, in the same year, only 42% thought there was a broad scientific consensus on climate change's reality (Leiserowitz et al. 2013). For a recent philosophical treatment of this issue of "balance" in news and scientific reporting, see Gerken (2020).

24 The apocryphal P. T. Barnum quote, "there's no such thing as bad publicity," is somewhat relevant here. Cf. "rage farming" or "rage baiting" (Schwemmer 2021; Jong-Fast 2022). Roughly, this is being controversial or outrageous in order to get (primarily online) attention. In some DINNER cases, this is plausibly an in-person analogue of such rage elicitation.

25 See Gerken's *Scientific Testimony*, especially 149–151. He calls this the *salient alternative effect*: "Roughly, this is people's disinclination to accept ascriptions of knowledge in the face of contextually salient error-possibilities." 149–150. See also David Lewis (1996)'s "Elusive Knowledge" for further parallels between the making salient of error possibilities such as climate change being a hoax which potentially causes a loss of knowledge in climate change, and making salient evil demon scenarios, which potentially causes a loss of knowledge in the external world. Worsnip (2021) also makes similar claims.

in all these cases in that it still somewhat increases the likelihood of begetting false beliefs or corrupting knowledge in the epistemic environment. Plausibly, the mere fact that something is being discussed can make the discussion appear legitimate and normalise it further.

Of course, even if I am right about this, one might reply that the *positive* epistemic effects garnered by Erica's victory will still outweigh any of the potential negative epistemic effects proposed above—if they even materialise. This could well be the case, but I do not think it is *immediately* obvious. If we assume that the other colleagues had *no beliefs* or were suspended about climate change, then it does look like there would plausibly be a sizeable gain in the inculcation of true belief or knowledge. Nonetheless, as I already noted above, following general demographics (from Tranter and Booth (2015) and Ballew et al. (2019)), the majority of the bystanders already *truly believe* or *know* that climate change is real. In light of this then, a natural question is the following: what positive epistemic effects *are* realised? Perhaps some bystanders achieve certainty? This seems unlikely; witnessing a creationist being dialectically trounced by a physicist does not strike me as an event that would make certain the reality of the Big Bang to someone who antecedently believes in it. I think a more plausible contention is DINNER** resulting in some small increases in confidence—be that either in the belief itself or in the evidence supporting it. And, of course, we must also allow that some (or all) of the deniers jettison their false belief and perhaps further replace it with a true belief about climate change. Granting all this, does it outweigh the potential legitimising negative effects I outlined above? I have my doubts, but admittedly lack a definitive answer to this. Either way, I am happy to concede for now that there are indeed epistemic benefits from Erica's engagement triumph that outweigh any possible negative ones. Even allowing this, however, I will now argue that it comes of little comfort for those of us potentially engaging with *controversial false assertions*.

3.2.4 Stocktake

So, where are we? I began by noting a lacuna in the engaging with false assertions literature related to the potential negative epistemic effects *outwith* the engager and the engaged. I connected this to another underappreciated aspect of the literature: namely, the myriad of ways an engagement can go. From there, I then looked at three different outcomes from an engagement—a loss from the champion of truth, an impasse, and a win. I then argued that each of these three options are not made equal insofar as a loss is far more negative epistemically speaking than a win is positive. Moreover, a stalemate is not a neutral outcome as the name might suggest but rather is better considered a win for the false assertor (with the caveat that it is a less decisive win than Bob's in DINNER). Clearly, none of this is particularly good news—-if any readers are harbouring Millian sympathies about the marketplace of ideas, the

arguments above suggest some scepticism to say the least.

With all this in mind, let us throw one more plausible idea into the mix: there is no real way for a prospective engager to know how an engagement will go beforehand. Therefore, pre-engagement, the epistemic risks from doing so look to far outweigh the potential rewards. Consider the following expected value calculations:

Outcome	**Probability**	**Epistemic (Dis)value**
Loss	0.2	-10
Stalemate	0.7	-5
Win	0.1	3

Table 3.2: *Expected epistemic value of engaging with controversial false assertions*

The expected epistemic value here works out at -5.2, meaning that the balance is *well* in the negative on average when we opt to engage with the sorts of assertions I have been discussing throughout. Of course, the numbers given above are somewhat arbitrary but they do track with the arguments laid out in the previous subsections. We already covered in §3.2.1 how disastrous a loss would likely prove to be in terms of the knowledge lost and/or justification defeated from the bystanders who previously believed in the reality of climate change, and the false belief inculcated in those who were suspended on the question, so there are good grounds for a weighty disvalue sum far more negative than the win is positive. I also argued that DINNER*, the stalemate situation, while clearly not as pernicious as the loss scenario, still looked to overwhelmingly favour the climate change denier, Derek, due to the way justificatory defeat would be apportioned among bystanders, thus it is also more epistemically concerning than the win is decisive. In fact, giving the win a plus-value as I have done is potentially somewhat charitable given the legitimisation worries so the expected value is plausibly even worse than given above. The probabilities stem from first the communications study cited earlier that suggested that 71% of such discussions end at an impasse (Cionea, Piercy, and Carpenter 2017), and second just general intuition that stalemate is the most likely scenario. I then put losing at nominally more likely due to the data that shows climate change scepticism *increased* with greater scientific literacy among Republicans (Funk and Kennedy 2020), but even if we say the two are equally likely *or* winning having a slightly higher probability (which I am confident would be a mistake), the expected epistemic value is still majorly weighted in the negative.

In fact, even supposing we put together a hugely *optimistic* calculation, the expected value remains negative:

Outcome	Probability	Epistemic (Dis)value
Loss	0.33	-3
Stalemate	0.33	-2
Win	0.33	3

Table 3.3: *Optimistic expected epistemic value of engaging with controversial false assertions*

Here we get -0.63 as the overall expected epistemic value which is evidently rather close to neutral, but we should be clear here that the set-up in these cases were supposed to be essentially "ideal" scenarios according to the philosophers in the extant literature (viz., closed-minded engagers, sans any epistemic injustices). Moreover, from what I argued above, the numbers in this table are rather implausible; there is nothing to suggest that each outcome is equally likely, and I argued that the negative effects in the losing case are far greater than the positives in the winning case. So, putting together an unrealistically optimistic scenario *still* works out as an overall negative, even when we eliminate the concerns that have taken precedence in the extant literature!

In sum, there are clearly far more dimensions and elements to consider when it comes to engaging with these sorts of *controversial false assertions* than has previously been appreciated by those working on this area, and at this point I think we have good reason to be markedly more reticent about objecting or engaging and any obligations or duties therein.

3.2.5 Online

At the beginning of §3.2, I noted that it is a virtue of the DINNER cases that they are relevantly analogous to possible online settings insofar as there are two interlocutors and an audience. The reason why I view this as a boon for my argument is because an arena where *controversial false assertions* are certainly prevalent is, of course, online—in particular, on social media sites (Vosoughi, Roy, and Aral 2018; Y. Wang et al. 2019; Suarez-Lledo and Alvarez-Galvez 2021). Readers themselves may even consider how many times they have observed or even participated in such political or medical arguments on, for instance, Facebook or Twitter.

Importantly, then, I propose that the arguments above about the epistemic dangers for bystanders *are* relevant to such online scenarios. Moreover, not only do they apply *mutatis mutandis*, but I will now argue that the concerns are actually *exacerbated* online compared to in-person. Specifically, and staying in lockstep with what has been discussed so far in this chapter, the focus in this subsection is the (even greater) difficulty one is faced with in securing the *winning scenario* (that is, DINNER**) in an online setting. Nevertheless, it is also important to note that the entire online situation

is often worse due to the platforms themselves being designed and indeed optimised in a way that amplifies hateful, incendiary, and violent content (Munn 2020).[26]

Let us consider one final vignette plausibly more common than any of the DINNER cases outlined above:

> SOCIAL MEDIA
> Greta is scrolling on social media when she comes across a long thread purporting to show that the Holocaust is a hoax. Greta knows this is false and so replies to the post, objecting to its claims. The author of the thread never responds to her, however. In fact, the only tangible effects Greta's response really had was that she amplified engagement on the post (she increased the comment counter and caused some of her own social media followers to click on the thread).

Earlier in §3.2.3, we granted that if an engagement is demonstrably successful inasmuch as the objector seems to strongly prove the false assertor wrong then it is plausible that it will have an overall positive epistemic impact. On top of this, it looked like the *only* scenario (of the three considered, at least) that could ensure a net positive. The problem here in SOCIAL MEDIA is that whether this is true or not, the normative pressure for the false assertor to respond to an engagement or objection does not exist in the same way that it does in-person. Therefore, even if we stipulated that Greta *would* be very successful if the Holocaust denier were to respond, the assertor can simply ignore her, and thus entirely prevent the possibility of this kind of *dialogical, engagement* victory for her. Crucially, this does not reflect as poorly on the denier the same way it would if the discussion were in-person.

To explain further, suppose SOCIAL MEDIA were face-to-face a la the DINNER cases. Were Greta to there engage with the denier and they were to simply shut down and make no response, this would look much like an engagement victory for Greta—i.e., her immediate objection or response was so effective that her interlocutor had no rejoinder whatsoever and just stayed silent. Thus, this is plausibly a winning situation and so a scenario where overall positive epistemic effects are realised.

However, the same reasoning does not seem to apply in the online version. When the Holocaust denier does not respond to an objection, an equivalent inference of their inability to respond does not seem to be as legitimately drawn because of the distinct dearth of normative pressure to engage with objectors online when compared with in-person, therefore they plausibly do not even have to take the *risk* of being dialectically trumped by anyone. Of course, a lack of response could be interpreted as the denier having no good answer for Greta, but at the very least it will not always be interpreted so—and I would contend not even *mostly*. Thus, when combined with the fact that an engagement victory is already the most unlikely scenario granted that one even gets

26 My thanks to an anonymous referee for suggesting I note this.

the *opportunity* for a dialogue, there are legitimate worries that the good scenario of DINNER** will be extremely rare online. And so, a regular outcome from this sort of engagement is just exposing the thread to more people,[27] and it is difficult to see how this could possibly be overall more epistemically beneficial than deleterious.

Neither does it seem like we can just ignore all such *controversial false assertions* online, however.[28] Were everyone who knows the falsity of such assertions to simply turn a blind eye and scroll past when they come across them, it would obviously be rather problematic. The most abhorrent of claims would only ever be accompanied by agreement—it is facile to predict negative epistemic consequences from this. Hence, *controversial false assertions* are plausibly even more problematic online than in any other arena, and their commonality means there is not even the comfort one might have had earlier that these situations do not arise particularly often.

3.3 UPSHOTS

The foregoing conclusions have all been rather pessimistic. I have suggested that there is much greater epistemic danger with engaging than has been appreciated thus far in the literature, and so have called for caution, even when the worries that other scholars have pointed out are accounted for. In this final section, then, I will employ these conclusions to yield some upshots—both theoretical and practical. The theoretical upshots will mainly focus on the problems my arguments might generate for certain philosophers who have considered obligations to engage or object in consequentialist terms; while the practical upshots will sketch some ways of potentially engaging to avoid the problems outlined above, and also suggest that some epistemically paternalistic practices may be the way forward.

3.3.1 Theoretical

Throughout, I have been particularly concerned with the consequences (and potential or expected consequences) from engaging with or objecting to *controversial false assertions*. I have shown that, on balance, the epistemic outcomes are broadly negative. Therefore, anyone who grounds any obligations or duties (or even permissions) to engage or object in *good epistemic consequences* from doing so looks to be in some trouble. The natural question then is: does anyone ground such normativity in consequences?

McKenna (2023), for one, explicitly appeals to an (epistemic) consequentialist framework when discussing obligations to engage with challenges to your beliefs.[29] McIn-

27 Cf. Saul (2021, 147–148). She calls this "amplification".

28 McIntyre (2020) also says much the same: "providing no response to misinformation was the worst thing one could do; with no rebuttal message, subjects were most likely to be swayed toward false beliefs." (220) He draws on Schmid and Betsch (2019) for this conclusion.

29 Specifically, see McKenna (2023, 105).

tyre (2018, 2020, 2021) also appeals to consequentialism when putting forward arguments as to why we ought to engage with *controversial false assertions* (although, of course, he does not use this term). Unsurprisingly, so does one of the original progenitors of consequentialism, Mill (1859). The rest (that took centre stage in §3.1) do not explicitly state precisely how they are thinking about this exact area, but their discussions are replete with allusions and references to consequences and outcomes so my arguments throughout are certainly ones that they ought to take seriously.

Lackey (2020b) is presumably the only one immune to these arguments due to the fact that her epistemic duty to object is a deontic one, so the consequences are immaterial. Nevertheless, in later work when discussing this duty, she appeals to a Singer-esque umbrella principle[30] which certainly appears consequentialist in nature, so it is possible that even she must take the arguments in this paper into consideration.[31]

In any case, all the authors highlighted in the introduction are to varying degrees interested in consequences and, perhaps implicitly, believed that aside from a few potential problems an objector can have when engaging, there was not much else to be overly concerned about. My arguments have shown that this is not true, and there are certainly further potential areas of epistemic danger outwith the pure dichotomy of the engager and their interlocutor.

3.3.2 Practical

Finally, let us turn to the practical upshots from the arguments and conclusions drawn throughout. First, for the online side of the engagement problems I have outlined, while it looks like there may be concerns with *individual* responses to *controversial false assertions*, there seems to be obvious, *institutional* responses that are in fact already implemented in the real world: epistemically paternalistic policies of deplatforming and censorship.[32] On social media, if one spreads *controversial false assertions* like climate change denial, their posts will be removed, and if they continue to do so repeatedly, they will be deplatformed (indefinitely banned from the site).[33] So, the thought goes, this sort of solution eliminates the engaging and objecting worries I have outlined throughout by simply cutting the problem off at the root; if people do not even *en-*

30 See, for instance, Singer (1972).

31 She calls this epistemic umbrella principle *Interpersonal Epistemic Duties*: "If it is in our power to prevent something epistemically bad from happening through very little effort on our part, we ought, epistemically, to do it." (Lackey 2020a, 287) While I would not say this is necessarily consequentialist, it can definitely be read as so (and its Singer origins gives some grounds for viewing it this way as well).

32 For the key works on epistemic paternalism, see Goldman (1991) and Ahlstrom-Vij (2013). I will discuss epistemic paternalism specifically connected to deplatforming and censorship (and the potential problems therein) in Chapter 4.

33 This was true in the past at least but since new management took over Twitter, this is no longer the case. Meta (viz., Facebook and Instagram) also seems to be following suit in some respects (McMahon, Kleinman, and Subramanian 2025).

counter such *controversial false assertions*, no one has to engage with them or object to them as there is nothing *to* engage with or object to, and so people's epistemologies cannot be adversely affected by them. Moreover, empirical data strongly supports this claim that such deplatforming and censoring practices are extremely effective in reducing the reach of bad actors and their assertions (Rauchfleisch and Kaiser 2021; Innes and Innes 2021; Jhaver et al. 2021).

Nevertheless, this is not a perfect solution. For a start, there are basic worries about granting large corporations and governments such epistemically paternalistic powers as the scope for abuse is obviously troubling (Aird 2023; Goldman 1991, 127). Second, even if implemented, such policies are not infallible in that some *controversial false assertions* will slip through the cracks (for a time, at least). In such cases, the problems outlined in this chapter will arise again in that it is a live question of what one should do in cases where one comes across a *controversial false assertion* that is yet to be removed. Third, there are historical examples of (attempted) censorship of books or films simply making them *more* popular such as *Lady Chatterly's Lover* (Baksi 2019), or *A Clockwork Orange* (Brew 2019). Fourth, there is some empirical work that suggests that deplatforming may increase polarisation and further entrench people in the views of the deplatformed individual as they follow them to a different site with more lax moderation rules and thus fall deeper down the rabbit hole and into an even more dangerous echo chamber (Ali et al. 2021).[34] Finally, even if we were to grant that such epistemically paternalistic policies *are* wholly effective in our online scenarios, this gives us no solution to our in-person DINNER cases insofar as it is not immediately obvious if we could "deplatform" Bob, Derek, or Frank from the dinner table in any meaningful way.

The final point is an interesting one, and worth exploring. After all, despite the potential problems with deplatforming and censorship above, it does look like a far more effective strategy in general for combatting *controversial false assertions* than simply engaging and disagreeing.[35] Are there in-person analogues for deplatforming that we could apply to DINNER and similar cases? Prima facie, it does not seem particularly plausible, but, in formal settings, they *do* exist. For instance, think about the practice of "striking from the record" in courtrooms. The Federal Rules of Civil Procedure state: "The court may strike from a pleading an insufficient defense or any redundant, immaterial, impertinent, or scandalous matter." (United States Courts) This is essentially the deplatforming or censoring of certain ideas, and, say, climate change denial presumably comes under "redundant", "impertinent", or

34 For the seminal philosophical work on echo chambers, see Nguyen (2020). For some more recent work, see van Oosterum (2025).

35 Particularly if one thinks that they will lose the engagement a la DINNER or DINNER*. Interestingly, Paglieri (2013, sec. 3) suggests that *likelihood of winning* is the main predictor for prospective engagers committing to arguing or not, so it is plausible that speakers are capable of identifying their prospects of success with some reliability. Thus, we have reason to think that finding strategies to employ when one believes they will be unsuccessful in the dialogue would be useful.

"scandalous". Again, however, this is a practice in a formal setting, so exactly how to apply it in our DINNER cases is not especially clear (Alex putting forward a motion to strike after Bob's espousals does not sound like a very effective strategy).

Nonetheless, there is something in the idea of striking from the record that is potentially of utility: the explicit recommendation that a statement or assertion is so outrageous that it does not deserve any consideration or response; in fact, we should expunge it entirely and move on as though it does not exist. Capturing this phenomenon in our informal cases is no mean feat, and likely beyond the scope of this paper, but I think a potential method of that is of *interrupting* or *closing the conversation*. For instance, Bob begins stating his climate change denial and Alex immediately interrupts and moves the conversation onto something else. This is clearly different from the serious engagement of objecting and arguing prevalent throughout this chapter and arguably avoids our engagement worries as Bob's assertion is cut short before it is added to the conversational record and properly discussed. Of course, there are again caveats here in that Bob might just not allow himself to be interrupted or the conversation to be closed and then we seem to be back to where we started with objecting, arguing, and so on, but as a first gloss these conversation closers have encouraging attributes.

There is a further strategy that I think is fairly obvious in light of the discussion throughout. This chapter has mainly focused on the effects on those outwith the engager and the engaged—the third-party observers. Thus, if we can remove this audience, then our only worries are the ones that have been discussed in the extant literature already (for instance, avoiding open-mindedness so as not to put your own knowledge up for grabs, or preventing any epistemic injustices). Therefore, if one of these sorts of discussions about *controversial false assertions* appears unavoidable, a good strategy to avoid the problems I have discussed is to try to have the conversation in a one-on-one setting.[36] In such a case, there is no worry about deleterious effects on third-party observers—because there simply are none. Of course, somehow ensuring the discussion takes place sans bystanders is not necessarily an easy task, but it is an important factor to keep in mind.

A final alternative that strays away from the sort of dialogical, pure engagement of objecting and arguing with someone—essentially, telling the person that they are wrong—that characterised the DINNER cases above is instead to employ narratives (Whitmarsh and Corner 2017) and avoid using certain terms when discussing the *controversial false assertion* at hand. For example, Arbuckle et al. (2014) found that, "emphasising terminology and narratives that focus on adaptation to weather variability rather than climate change may be better received and more effective when working with farmers." (515) So, the idea is that the *reality* of the matter is still under discussion (i.e., genuine extreme weather events caused by climate change) but

36 This is a strategy also recommended by McCormick (2023).

it is couched in terms that avoid the charged, politicised language of, say, "climate change" or "global warming", and is not phrased or framed in a way that is necessarily *objecting to* or even *disagreeing with* the engaged. Not only does the aforementioned empirical data suggest that this is an effective engagement strategy simpliciter, the relevance here is that this kind of engaging plausibly avoids the worries outlined in this paper. This is because the engager can (hopefully) avoid getting into an argument like the protagonists in the DINNER cases, and thus this more cautious engaging should not carry with it the same possibility of dialectical failure or stalemate that plagued those examples.[37]

The practical suggestions above are mere sketches, and the exact effectiveness of them is not something I can guarantee. Nevertheless, as a starting point to potentially avoid the engagement issues I have discussed throughout, I think these show some promise, and ought to be explored further.

3.4 CONCLUSION

I began with a broad overview of the current literature on engaging with challenges to our beliefs, noting a lacuna relating to the possible epistemic effects outwith merely the engager and the engaged. I then went on to discuss three versions of the same case, each with a different outcome, where the highly pernicious nature of certain engagements was shown. Overall, I proposed that the possible effects on third-party observers when it comes to engaging with *controversial false assertions* are so wide-ranging that they swamp any of the benefits previously discussed in the literature. Alone, these are novel arguments and conclusions but I drew on them to suggest that this should give pause, and make us reconsider many of the obligations and duties to object or engage put forward thus far. Finally, I closed with some suggestions of how to possibly avoid these engagement worries that plagued vanilla engagement and objecting, looking at some epistemically paternalistic policies among other, more individualistic, responses.

37 On the other hand, there is considerable scepticism in science communication about this idea of simply presenting facts to the public being a successful engagement strategy for understanding science–for more on the deficit model and its flaws, see, among many others, Miller (2001), Nisbet and Scheufele (2009), and Simis et al. (2016). Nevertheless, these issues are more focused on science communicator/layperson interactions which are at least somewhat different from the more ordinary peer-to-peer discussions that I have been thinking about here. For a philosophical treatment and scepticism that these issues are even ones of information deficit, also see Levy (2021, 24–27). My thanks to an anonymous reviewer for urging me to discuss this.

Linking Interlude III

In chapter 3 I argued that objecting can be an epistemically dangerous business, and we potentially ought to be more reticent than has previously been appreciated. Nonetheless, the problems of *controversial false assertions* (people believing, asserting, spreading them) remain. Hence, my sketching of some solutions towards the end. One of those discussed was epistemic paternalism—in particular, policies of censoring or removing people from platforms. Prima facie, this might look like a rather promising strategy: after all, how can such misinformation adversely affect people if they simply do not come across it? In the forthcoming chapter, I argue that while epistemic paternalism may make for an effective tool in tackling misinformation, the risks incurred by employing such a weapon are too great.

Chapter 4

A puzzle of epistemic paternalism

In recent years, misinformation, fake news and conspiracy theories have abounded, plausibly having the consequences of multiple deleterious events such as drastically affecting global health measures to oppose the Covid-19 pandemic. In response, different strategies have been proposed to combat such misinformation, fake news, and conspiracy theories. One suggested approach has been that of epistemic paternalism—viz., non-consultative interference in agents' inquiries for their own epistemic improvement. While the extant literature on epistemic paternalism has mainly discussed whether it is (ever) justified to interfere with others in this way, here in this chapter, I primarily focus on the potential implementation of widespread epistemically paternalistic policies (such as no-platforming and censorship) and the consequences therein. I argue that pursuing epistemically paternalistic policies to combat such misinformation, fake news, and conspiracy theories leads to a hitherto unnoticed puzzle for proponents of epistemic paternalism. Central to the puzzle is the idea that those (e.g., governments, corporations, social media giants) who actually can—i.e., have the requisite power to—enact widespread epistemically paternalistic policies seem the institutions who are least suited to having such informational control over the populace. Thus, epistemic paternalism (in this context) appears a sword without a hilt; while it may prove an effective strategy in tackling misinformation, fake news, and conspiracy theories, there is no way to use it without incurring serious risks.

4.1 Introduction

Consider the following common assumption:

> Infodemic Folk Theory
> Misinformation, fake news, and conspiracy theories are crucial contemporary problems, more so than ever before.

This thought is borne out in a variety of ways; in public health, the WHO's director-general said of Covid-19: "We're not just fighting a pandemic; we're fighting an info-

demic," (The Lancet Infectious Diseases 2020); in the media, headlines read: "You're living in the Golden Age of Conspiracy Theories" (Stanton 2020); in the populace's general conscious, with 83% of Europeans and 68% of Americans seeing made-up information/fake news as key issues facing democracy today (European Commission 2018; Mitchell et al. 2019); and in prevalent, concrete issues such as the UK riots in 2024 (Chantler-Hicks 2024), or measles outbreaks in Texas (Halpert 2025).

Naturally, due to the general concern of people believing lots of false and dangerous things, not to mention the myriad negative (epistemic and otherwise) consequences that seem to stem from such beliefs, researchers have wondered what strategies can be used to effectively tackle such conspiracy theories and misinformation. One such approach philosophers have recently considered (for instance, Castro, Pham, and Rubel (2020) or Brown (2021)) is that of *epistemic paternalism*—that is, non-consultative interference in agents' inquiries for their epistemic improvement (Ahlstrom-Vij 2013). If agents can be shielded from such misinformation, the thought goes, then it will not be able to adversely affect them—plausibly, say, protecting them from coming to hold the relevant false beliefs or stop them from having (the justification of) their knowledge defeated.[1] In this chapter, however, I argue that pursuing epistemically paternalistic procedures in response to the problems of the INFODEMIC FOLK THEORY leads to a puzzle previously unnoticed in the epistemic paternalism literature and by its proponents. Central to the puzzle is the idea that the corporations, social media giants, and governments who actually *can* enact widespread epistemically paternalistic policies of removing from platforms repeat offenders of spreading misinformation or censoring individual instances of fake news or conspiracy theories are those very institutions that we would not want to grant the power of informational control over the populace.

Here is how this chapter will proceed: I begin in §4.2 by defining (and, indeed, *not* defining) some key terms, giving a broad overview of misinformation, fake news, and conspiracy theories before homing in on the conception of epistemic paternalism I will employ and showing how classic online content moderation strategies are instances of the phenomenon. In §4.3, I give the puzzle of epistemic paternalism in full, justifying each of its premises individually, followed by locating the significance of having discovered the puzzle. I then turn in §4.4 to considering some objections and responses to the puzzle. Nevertheless, I find that none of the objections prove particularly problematic and, thus, the puzzle stands. In §4.5, I wrap up with some concluding remarks.

1 Simion (2024a) gives a taxonomy of five different ways mis/disinformation can generate ignorance: (1) generating false belief, (2) misleading defeat, (3) inducing epistemic anxiety, (4) confidence defeating, (5) exploiting pragmatic phenomena (1214). While she notes this is potentially not exhaustive, it is an excellent basis for the ways in which misinformation can adversely affect one's epistemology.

4.2 PUTTING PIECES IN PLACE

4.2.1 Misinformation, fake news, and conspiracy theories

To be clear, I will not be perfectly defining misinformation, fake news, nor conspiracy theories in this set-up subsection—those are essentially book-length (or at least chapter-length) discussions in their own right.[2] Nor is it necessary to have such conceptual analysis in hand for the argument in this chapter. Nevertheless, *some idea* of the key terms is important so I will home in on some simple examples and fairly unobjectionable stipulations.

For the rest of this chapter, I will use "misinformation" as a suitable umbrella term, meaning something like information primarily intended to mislead agents into developing false beliefs.[3] I take misinformation to include under it fake news and conspiracy theories, and anything else we might think is appropriately intentionally misleading—such as deepfakes.[4]

To give some helpful examples most will be familiar with, we might think that there was general *misinformation* surrounding the Covid-19 pandemic. Specific examples being the conspiracy theory that 5G telecommunication towers spread the virus—which in turn lead to the vandalism and destruction of such towers along with harassment of telecom personnel (Satariano and Alba 2020); or fake news where data from the *Vaccine Adverse Event Reporting System* (VAERS) apparently showed that vaccines were extremely dangerous—this was misleading for a number of reasons but primarily because the data (while plausibly having the *appearance* of being extremely dangerous to uninformed members of the laity) actually showed extremely low percentage of adverse effects (see Lyons (2021) for more). A different—yet plausibly most crucial example—is that of factually *accurate* but misleading content as discussed in Allen, Watts, and Rand (2024) which they found to be "46-fold more consequential for driving vaccine hesitancy than flagged misinformation." (1) One of their key examples is that of a (true) headline which read: "A healthy doctor died two weeks after getting a COVID vaccine; CDC is investigating why", the thought being that while this is accurate reporting, its general phrasing, implicatures, and so on, lends itself to false belief generation—about the safety of vaccines, presumably. Clearly, however, this is neither fake news nor a conspiracy theory.

These are the sorts of misinformation I will be most interested in in this chapter.

2 For misinformation, see Harris (2024); for fake news, see Bernecker, Flowerree, and Grundmann (2021) for a collection on the epistemology of fake news, and Gelfert (2018), Fallis and Mathiesen (2019), and Grundmann (2023) for some attempts at defining the phenomenon; for conspiracy theories, see Chapter 5 where I will give a full overview of the literature and eventually offer my own account.

3 Although again see footnote 1.

4 These have received a good amount of philosophical discussion from the seminal pessimistic work from Rini (2020) to more scepticism about their apparent widespread epistemic dangers by Harris (2021) and Habgood-Coote (2023).

Precisely what each of these are is not overly important so long as we can agree that these things can and do exist, have caused some clear epistemic (and beyond) problems, and there are some unobjectionable central examples. The idea is that so long as we have these, we can discuss some potential measures we can use to tackle them, and from there my puzzle of epistemic paternalism will be generated.

4.2.2 Epistemic paternalism

Let us now turn to laying out epistemic paternalism. Following orthodoxy in the literature,[5] I will here employ Ahlstrom-Vij's definition which has three jointly necessary and sufficient conditions:

> [A] practice is epistemically paternalistic if and only if it interferes with the freedom of inquirers to conduct inquiry in whatever way they see fit (*the interference condition*) without consulting those interfered with on the issue of whether they should be interfered with in the relevant manner (*the non-consultation condition*), and moreover interferes—exclusively or not—for the purpose of making those interfered with epistemically better off (*the improvement condition*). (Ahlstrom-Vij 2013, 61)[6]

The first condition relates to an agent's freedom to collate, evaluate, and discover evidence being impeded by some interferer; the second to the fact that such interferences with these processes are done without their approval (this is also where the general suspicion of epistemic paternalism stems from); the third is what makes it an interesting discussion—despite the first two seeming prima facie bad or inappropriate, it is all in fact done for the agent's *improvement*. Goldman's classic example of epistemic paternalism stems from the Federal Rules of Evidence for United States Courts and Magistrates, where, "evidence of previous crimes by the accused is not admissible to help prove that he committed the present crime." (Goldman 1991, 116) However, the reason this (accurate and plausibly relevant) evidence is withheld is so as not to prejudice or mislead the jury about *the case at hand*—in other words, as Goldman notes, this is an *epistemic rationale* insofar as this rule seems to be in place to help jurors come to true beliefs about the case. We can see how this case meets our three conditions: the jury's inquiry is interfered with insofar as they do not have access to all the evidence

5 See, for instance, among many more, Bullock (2018), Croce (2018, 2020), Broncano-Berrocal (2020), Godden (2020), Jackson (2020), Simpson (2021), and Kitsik (2023). Note also that the seminal work from Goldman (1991), as we will soon see, contains a merely less precise version of the forthcoming definition from Ahlstrom-Vij: "I shall think of communication controllers as exercising epistemic paternalism whenever they interpose their own judgment rather than allow the audience to exercise theirs (all with an eye to the audience's epistemic prospects)." (119)

6 Following Jackson (2021, 135) I actually prefer a *non-consensual condition* over the non-consultation condition above for the reason that a prospective paternaliser *could* consult someone, and then simply interfere against their wishes anyway. Intuitively, this is still a paternalistic act, thus a non-consent condition is better. Nonetheless, I will set this aside here as it has no bearing on the discussion to come.

that may be relevant; they were not consulted on this interference as it is determined by the judge and the Federal Rules of Evidence; but this is all done so that the jury (hopefully) comes to a true belief about the case—viz., it is done to make the jury epistemically better off.

4.2.3 Censorship and no-platforming as epistemic paternalism

Epistemic paternalism can come in a variety of forms,[7] but we will only focus on two here for this chapter: the practices of *no-platforming*[8] and *censorship*. Peters and Nottelmann (2021) understand no-platforming as follows: "the practice of denying someone the opportunity to express their opinion at certain venues because of the perceived abhorrent or misguided nature of their view(s)." (7231) We can broadly adopt that definition here, where "certain venues" is mainly going to refer to social media spaces but could also include, say, campuses or television.[9] For censorship, we can employ Joshi's recent definition: "censorship involves the intentional act of either preventing some claim from being successfully communicated or significantly disincentivizing such communication." (Joshi 2024, 2) Much like the above discussion of misinformation, precise definitions will not bear too heavily on the discussion here: an intuitive understanding of the central pieces is sufficient. Accordingly, then, we can briefly discuss an example to show how such practices are indeed epistemically paternalistic:

> MISINFORMATION MARK
> Mark used to run a popular Facebook "news" page called *Coronavirus Truth Org*. Despite the name, Mark would almost exclusively post articles containing or entirely made up by misinformation, fake news, and conspiracy theories about Covid-19 and the pandemic as a whole. In fact, Mark's site was an early proponent of the conspiracy theory that 5G towers spread the virus. These individual posts were initially removed (i.e., *censored*). However, *Coronavirus Truth Org* continually perpetuated misinformation until the full page was no-platformed from Facebook and all posts deleted.

We can now go through and see how the above actions (both the instance of censorship and no-platforming) meet the three conditions and so constitute acts of epistemic paternalism. First, the interference condition. Any agent who wished to, say, inquire

7 Sometimes to the detriment of the phenomenon's plausibility; Jackson (2021, 135–137) discusses how writing a philosophy book and triaging what discussions and views one contains within it could plausibly be countenanced as epistemic paternalism (which does not seem to be in the spirit of the discussion in the literature). This leads her to propose a "*significant* interference" modification.

8 Outwith philosophy, the term mainly used is "deplatforming". As far as I can tell, both refer to the same thing. In any case, I will use "no-platforming" throughout.

9 For the classic discussion of no-platforming related to campuses and "academic freedom", see Simpson and Srinivasan (2018). For a more recent discussion, see Elford (2023).

into whether there is any link between 5G and Covid-19 now have had their inquiry interfered with in that there is some evidence (*Coronavirus Truth Org*'s articles) that they no longer have any access to.[10] Second, the non-consultation condition. Presumably, not every single person in the world (if any) who had or might have been interested in investigating 5G and Covid-19 relations were consulted on the initial censorship nor the no-platforming.[11] Finally, the improvement condition. The removal of the articles and subsequently the page were clearly done for agent's epistemic improvement in that it was to prevent them from acquiring misleading evidence, which in turn sought to prevent them from coming to hold false beliefs. Thus, both censorship and no-platforming can be instances of epistemically paternalistic acts—and these sorts of acts are the ones central to this chapter. Now that all the pieces are in place, we can turn to the puzzle.

4.3 THE PUZZLE

In this section, I give the puzzle in full before going through and motivating each of its premises. I then consider the import of the puzzle and its consequences. First, however, it may seem like something salient to discuss is whether epistemic paternalism (of the sort here or indeed in general) is *ever* justified—just because it makes for a potential response or solution to a bad thing (in this case, misinformation) does not inherently justify it. Nevertheless, I am going to mainly set aside this question and assume that it is at least sometimes justified—namely, in the scenarios of misinformation and the like central to this chapter.[12] Instead, I home in more on issues surrounding the potential *implementation* of the sorts of no-platforming and censorship epistemic paternalism raised above. While it is a well discussed topic in the epistemic paternalism literature *who is justified* in acting epistemically paternalistically,[13] my focus is on the less discussed area of who *can* (i.e., who *has the power to*) enact widespread epistemically paternalistic policies (of no-platforming and/or censorship) which would affect vast numbers of people. The answer to this is a rather short list: institutions such as the government, and corporations such as social media

10 One might wonder whether eliminating misleading evidence of this sort really constitutes epistemic paternalism. I will later consider this objection in full in §4.4.2.

11 Moreover, if anyone *were* consulted and/or asked for the page or its posts to be removed, this clearly would not be epistemically paternalistic anyway.

12 For discussion on justifying epistemic paternalism, see, among others, Goldman (1991), McKenna (2020, 98–101), and Bondy (2020, 147–153). Additionally, while I am assuming that epistemic paternalism is at least sometimes justified (to get the puzzle off the ground), the puzzle coming shortly can actually contribute to the debate of epistemic paternalism's justificatory status in general. If the puzzle turns out to be true, then it is possible powerful organs of society would be unjustified in acting in widespread epistemically paternalistic ways. My thanks to an anonymous reviewer for pointing this out.

13 For instance, Goldman (1991) thinks *experts* (in a more social, credential driven respect) are justified to undertake epistemically paternalistic actions related to their area of expertise. Cf. Croce (2018) who argues against this and proposes *epistemic authorities* instead.

giants (Facebook, Youtube, Twitter) or Google. The puzzle arrives in that it seems these groups are (somewhat uniquely) the exact organisations we would not want to have the power of no-platforming or censorship. As Goldman writes:

> Epistemic paternalism on the part of isolated individuals is quite a different matter from paternalism exercised by the state, or any other powerful organ of society. There are historical reasons for being very cautious about state control of information. (Goldman 1991, 127)

For instance, proponents of epistemic paternalism often suggest science communicators as plausible candidates for justified acts of epistemic paternalism (John 2018; McKenna 2020; Medvecky 2020). John (2018) talks about a climate scientist making an "[epistemically] effective assertion" over an "honest assertion" (83–84) which is essentially non-consultative interference for agents' epistemic benefit. Nevertheless, this is not problematic for the puzzle to come despite my focus here in part on institutions—the thought being that science communicators are part of such institutions. In fact, I think the argument here complements nicely this extant literature.

To explain, suppose a science communicator acts in some epistemically paternalistic way when publishing a news article on vaccine safety; say, they cut some discussion about safety statistics because they fear—correctly—that it would prejudice the laity into getting the wrong idea about the safety of the vaccine. The issue I am pointing towards is that this (epistemically beneficial) article could be swamped (or at least surrounded) by misinformation and so has the potential to go unnoticed or have its possible benefits annulled. Therefore, we might reasonably seek ways to eradicate this surrounding misinformation—immediate methods of which will be widespread no-platforming or censorship. This is where the requisite power of implementation comes in, bringing with it a puzzle: only large institutions of the kind highlighted above can do this sort of widespread no-platforming or censorship—such control is not available to individual science communicators. However, these large institutions, as I will shortly argue, are not well-positioned to act reliably in agents' best epistemic interests.[14]

Now, let us lay out the puzzle explicitly:

I. There are pro tanto reasons to implement epistemically paternalistic policies to effectively combat misinformation.[15]

II. Those who enact such epistemically paternalistic policies should be well-positioned to act reliably in interferees' best epistemic interests.

III. The only plausible candidates for enacting such epistemically paternalistic

14 My thanks to an anonymous referee for raising this point about science communicators.

15 I am here taking "epistemically paternalistic policies" to mean formalised policies of no-platforming and censorship.

policies to effectively combat misinformation are institutions like large corporations or the government.

IV. Such institutions are not well-positioned to act reliably in interferees' best epistemic interests.

V. Therefore, epistemically paternalistic policies to combat misinformation cannot be enacted.

VI. Therefore, misinformation cannot be effectively combatted.

For a start, I take this argument to be intuitively and prima facie plausible. III, moreover, is straightforwardly true; a lone individual clearly lacks the requisite power to enact widespread epistemically paternalistic policies of no-platforming and censorship.[16] In fact, in some respects we might think a key strategy of theirs may be to petition these very powerful institutions. Nevertheless, the rest of the premises require greater motivation, so let us now turn there.

4.3.1 Premise I

As stated above in §4.2.3, the sorts of epistemically paternalistic acts I am interested in here are those of no-platforming and censorship. So, when it comes to utilising these in the name of combatting misinformation, for the former the outcomes we would generally see would be regular purveyors of misinformation having spreading privileges revoked—be that through social media bans, television and/or conference blacklisting, and so on. For the latter, it would primarily be the removal of individual pieces of misinformation such as the deletion of certain posts. This differs from a fully fledged no-platforming insofar as unknowing people could accidentally spread misinformation and see their posts removed without themselves being completely no-platforming a la MISINFORMATION MARK. The overall result of heavily implemented policies of this sort is that agents in such regulated epistemic environments are unlikely to even *encounter* such misinformation.

So, why might we think these make for effective strategies in combatting misinformation? The remainder of this subsection will provide evidence to answer this query.[17] Starting with the philosophical literature, Levy (2019) speaks of the dangers

16 This is not to say that individuals have *no power* whatsoever. Reporting misinformation or fact-checking acquaintances who are (inadvertently or not) repeating misinformation could potentially help with combatting misinformation but it seems unlikely that these strategies will be as immediately effective as a strong institutional response that affects entire populaces. I address this thought in more detail in §4.4.1.

17 Of course, these kinds of epistemically paternalistic acts are not the *sole* way of combatting misinformation. Rini (2017), for instance, suggests third-party fact checking agencies such as Snopes or FullFact.org could be useful for tackling misinformation. There is existing research that fact-check prompts on posts are effective at reducing misinformation's deleterious effects (Pennycook and Rand 2022)—although see Thorpe et al. (2022) and Tulin et al. (2024) for dissenting evidence. Either way,

of platforming problematic speakers in terms of the pernicious epistemic effects it can have on the wider epistemic environment:

> An offer of a platform is a signal that those who issue the invitation consider the person worthy of a respectful hearing. It is a signal that the inviters consider the speaker sufficiently expert, or sufficiently representative of expertise to have an opinion on that topic that should be taken into consideration. (Levy 2019, 495–496)[18]

So, should such a problematic speaker be no-platformed, we remove this worry of conferring legitimacy upon them and their illegitimate opinions (say, some brand of misinformation). Fantl (2018) discusses a similar idea:

> [I]f we know that the speakers… are uttering falsehoods, then we are prioritizing those other values over the value of truth because we are allowing falsehoods an inroad to the university that they wouldn't otherwise have. (200)

Fantl argues that this sort of epistemic harm[19] is impermissible and thinks the chief way to prevent this harm from occuring is by no-platforming such falsehood-uttering speakers; the extension of this reasoning to online scenarios is, I think, straightforward. Castro, Pham, and Rubel (2020) note the success of an epistemically paternalistic policy from Facebook which involved the demotion of fake news and suggest that, "The policy, if successful, will protect users from internalising attitudes that would be inauthentically held." (38) In fact, the main concern of most of the philosophical literature is whether the epistemic paternalism is *justified*—the effectiveness or benefits of such strategies are generally taken as a given.

Nonetheless, there is also an abundance of convincing empirical data on the effectiveness of such interventions in combatting misinformation. Chiou and Tucker (2018) write, "After Facebook's ban on advertising by fake news sites, the sharing of fake news articles fell by 75%." (1) There seems a very simple diagnosis here: people did not see the articles any more, so they did not share them. Shen and Rosé (2022)

my argument is not predicated on epistemic paternalism being the only way of combatting misinformation, merely that we have good reason to think it is a very effective one.

18 One might object that the real concern here is more about the reputation of the host party as opposed to any altruistic, paternalistic concern with the epistemic environment or epistemic well-being of others. Certainly, some acts of no-platforming or censorship could be done out of these more self-centred concerns (and no doubt there are actual examples of these), but if they are then they simply no longer qualify as epistemic paternalism anyway as they would not meet the improvement condition (the interference would not be for the purpose of making the interferee epistemically better off) and thus is not too relevant to the discussion here. This general objection, and the sort of category error worries that arise therein, is discussed in more detail in §4.4.3. My thanks to an anonymous reviewer for raising this point about reputational concerns.

19 He also discusses potential "psychological harm" (184–188) and "intrinsic harm" (189–197) that can be done to agents by platforming problematic speakers but these are separate from the epistemic concerns central here.

discuss the "quarantining" of two subreddits, /r/The_Donald and /r/ChapoTrapHouse on the website *Reddit*. Quarantining is a process that can be employed by the site that limits access to particular subreddits (i.e., smaller groups within the larger site) without wholly deleting them. While not as extreme as a full no-platforming, they found that it did reduce the number of new users and the popularity of posts. No access from anyone (i.e., a complete no-platforming) would obviously limit the new users and popularity even more. Rauchfleisch and Kaiser (2021) write in their study on no-platforming:

> Our analysis shows that [no-platforming] is effective in minimizing the reach of disinformation and extreme speech, as alternative platforms that will allow this kind of content cannot mitigate the negative effect of being [no-platformed] on YouTube. (1)

Innes and Innes (2021) look at the no-platforming of conspiracy theorists David Icke and Kate Shemirani—both of whom had extremely popular online presences prior to their removal from social media. This research concurs with the previous: no-platforming significantly limits the spread of misinformation (but with the caveat that it does not eliminate it entirely). Finally, Jhaver et al. (2021) take a similar line of research, looking at the no-platforming of infamous demagogues Alex Jones, Milo Yiannopoulos, and Owen Benjamin. In keeping with those surveyed so far, their research showed no-platforming to be a success in massively reducing the online impact of these individuals.[20]

On the whole, empirical data strongly suggests that no-platforming and censorship of bad actors on social media is extremely effective in reducing their reach and thus limiting the spread of misinformation—and the data for this is considerably more solid than any of the other strategies for combatting misinformation touched on like fact-checking. So, overall, between philosophical support and near-consensus empirical data, we have rather good reason to accept that such epistemically paternalistic policies implementing no-platforming and censorship would be an effective strategy for dealing with misinformation.

4.3.2 Premise II

II is a normative claim about how we would want those who would potentially enact epistemically paternalistic policies to operate. First, let us get clear on "best epistemic interests." In the epistemic paternalism literature, presumably down to Goldman's weighty influence, this is mainly understood on veritistic lines—that is, that true belief is paramount.[21] Here, this is entirely sufficient but there is no reason why we

20 And the fact that all three users subsequently came back to Twitter after new management took over is likely relevant data.

21 Although see Pritchard (2013) for a different conception of epistemic value directly connected to epistemic paternalism.

cannot take a broad pluralist stance where "best epistemic interests" could be anything from a knowledge-first approach to the promotion of epistemic virtues. Even a non-consequentialist, *respect for truth* framework (Sylvan 2020) would likely still work for my purposes here—nothing in the puzzle really hinges on a specific conception of epistemic value.

Next, let us turn to the "should" featuring in premise II. I am taking this to be a sort of role ought, where its denial would imply negligence.[22] Consider an example, inspired by Goldberg (2017), where we are talking about rubbish collection instead of epistemic paternalism and the second premise reads: *Those who collect rubbish should be well-positioned to collect rubbish reliably*. That is to say, they should possess the means to collect rubbish, be equipped with the ability to do so, have the opportunity to exercise the ability, and so on. If they purport to be occupying the rubbish collector role but fall short in any of these ways, then there is a (role-specific) normative defect here; they would be alleging to occupy a role that requires doing something that they are not *well-positioned* to do. It is in this normatively narrow, role-specific sense that rubbish collectors should be well-positioned to collect rubbish reliably.

This runs analogously with the roles in the puzzle at hand here. A precondition for enacting epistemically paternalistic policies appropriately is that those who enact them *should* be well-positioned to act reliably in the inteferees' best epistemic interests. If not, then they are failing in their role *as* paternalistic-policy-enactors. Of course, being well-positioned to act reliably in an interferee's best epistemic interests is clearly not necessary to act epistemically paternalistically tout court towards someone; imagine a systematically unreliable teacher who sincerely does not care about their students' epistemic status but nevertheless non-consultatively interferes in their inquiry some way to improve them epistemically. The key point rather is that these powers of no-platforming and censorship, while plausibly effective at dealing with misinformation, have massive scope to be used and abused beyond this. This premise acts as a sort of safety net in our wanting to try to prevent such abuses; at the very least those enacting such policies should be *well-positioned* to act reliably in interferees' epistemic interests. The problem, as stated in IV, is that these institutions are not well-positioned in this way.

4.3.3 Premise IV

IV is a descriptive claim about the reality of how these institutions are not well-positioned to act reliably in interferees' best epistemic interests. I suspect that such a claim enjoys significant intuitive support but supporting evidence can be offered

22 For discussion of such role oughts, see Feldman (2000, 2008). A different sense of "should" would perhaps be whether institutions should try to epistemically improve agents at all—i.e., is it their duty to help in this fashion? For excellent discussion that is somewhat related on various *should*s and social expectations, see Goldberg (2017, 2018). For a specific (although individualistic) discussion of our duties towards conspiracy theories and misinformation, see Terzian and Corbalán (2021).

regardless. For instance, a whistleblower report from inside Facebook revealed that higher-ups at the company knew that the content algorithm pushed users into further radicalisation and promoted the growth of QAnon but they deliberately elected to not ameliorate the issue (Gilbert 2021). A 2018 study found that false stories are 70% more likely to be retweeted on Twitter than true ones (Vosoughi, Roy, and Aral 2018). Plausibly, a social media site's main interest is keeping users on the site, and if misinformation, fake news, or conspiracy theories are instigating that, then rectifying such issues is unlikely to be a priority and thus they are not well-positioned to act reliably in agents' best epistemic interests. Additionally, these large institutions are companies that are constrained in how they *can* act—i.e., they have fiduciary responsibilities to shareholders. If (when) such responsibilities (e.g., keeping users on the site to generate advertising revenue) conflict with interferees' best epistemic interests, then the fiscal responsibilities will win out every time.[23]

Governments do not fare much better. The infamous Brexit bus, plastered with the false claims that the UK sends £350 million to the EU every week and instead this money could be used to fund the NHS is an obvious instance of politicians sacrificing its people's epistemic health for potential political gain (Dallison 2019). Or consider manipulated evidence from Western governments to justify the Iraq War (Doherty and Kiley 2023). Empirical evidence on governmental abuses of the public's epistemic health is widespread. Moreover, the two governments in question here are still *democratic* ones without ultimate control over the flow of information in the epistemic environment. Authoritarian governments that did (or do) have this power of non-consultative interference over the public certainly did not use it for the public's best epistemic interests. Consider Russia's current propaganda campaign regarding the war in Ukraine (Bond 2024) or the Chinese government's still-ongoing censorship of the Tiananmen Square Massacre (Griffiths 2019). Democratic or non-democratic, history is littered with governments from all over the world systematically not acting in their constituents' best epistemic interests. Thus, I propose they are not well-positioned to act reliably in interferees' best epistemic interests.

In sum then, with the premises justified, the puzzle of epistemic paternalism stands; we have an excellent strategy for combatting misinformation, but we should not use it because those solely with the power to do so, should not be given such power.

4.3.4 The import of the puzzle

While the general problem pointed toward in this paper (can those with the power to change things for the better actually be trusted to do so) is a familiar one, there is novel import in my identifying of this puzzle. We have seen that epistemic paternalism seems to be an effective strategy in tackling misinformation and, crucially,

23 As I noted all the way back in §1.1: the alethic pales in comparison to the dollar.

misinformation, fake news, and conspiracy theories are endemic in the epistemic environment—recall the INFODEMIC FOLK THEORY. Therefore, this culminates in good reason to pursue the implementation of epistemically paternalistic policies. The significance of my puzzle is that it points towards a problem that has been thus far overlooked by proponents of epistemic paternalism; namely that, irrespective of how effective epistemically paternalistic policies may be in combatting misinformation, such policies should not be enacted because the institutions who could enact these policies are not well-positioned to act reliably in the agents' best epistemic interests. But still my opponent might push back on this being of much interest. As a reductio, plausibly the best way to prevent murder is an extremely advanced and intrusive police state where everyone's actions are carefully tracked and monitored, and privacy is a thing of the past. Despite this probably being an effective strategy to combat murder, we would not want to use it for obvious reasons (breaches of human rights for one), but it does not seem that there is a "puzzle" here nor much of interest to discuss. So, the rejoinder might go, my puzzle is analogous.

Unsurprisingly, I do not agree. For a start, no matter how it works out in preventing murder (or any other crime for that matter), the police state has inherent problems from the off—e.g., the human rights abuses mentioned. The epistemic paternalism discussed in this paper only has *potential* problems from *potential* abuses that are directly tied to only those with the power to enact the policies in question. Were it *guaranteed* the powers of no-platforming and censorship were only used accurately, effectively and for unadulterated good (e.g., combatting misinformation, only no-platforming genuine bad actors, and so on) then presumably there would not be many worries about implementing such policies (aside from the personal sovereignty concerns familiar from extant discussions (Croce 2020)) which is substantively different from the police state. The problem—and puzzle—comes from the fact that we *cannot* guarantee such benign uses. In fact, because of how the only institutions with the power to enact such policies are positioned regarding people's epistemic interests, we can guarantee near the opposite. Nevertheless, these are still *contingent* worries, and this is part of what makes the puzzle of interest.

On the other hand, as some suggestive conjecture, there is a sense in which the puzzle of epistemic paternalism may prove somewhat inevitable. If we start with the reasonable assumption that there is some connection between untrustworthy institutions and the growth of conspiracy theories—i.e., when our institutions are untrustworthy, the more misinformation or conspiracy theories will abound.[24] In such a scenario, epistemically paternalistic policies would still be an effective method of combatting such misinformation but, because the untrustworthy institutions will not be well-positioned to act reliably in agents' epistemic interests, such strategies should

24 This is supported by empirical data. See, for instance, Pierre (2020), Pummerer et al. (2021), and Mari et al. (2021).

not be used. However, in a scenario where our institutions *are* trustworthy and so would potentially be well-positioned to act reliably in agents' best epistemic interests, there would plausibly be less conspiracism and misinformation, and so epistemically paternalistic policies may not even need to be used anyway. In a slogan: whenever epistemic paternalism may be needed, it cannot be used, and if it ever could be used, it is probably not needed.

4.4 RESPONSES

In this penultimate section, I will consider and reject three objections to various parts of the puzzle. I will not attempt to resolve the puzzle myself here in this chapter (nor in this thesis)—I instead am just looking to defend it from a few separate responses. I argue none prove fatal to the arguments above and so the puzzle stands.

4.4.1 Response 1

Can only *institutions* do this job of epistemic paternalising? Perhaps one could argue that this extremely strong and widespread institutional response to misinformation is unnecessary and that instead individuals can do the job of demoting, removing, and combatting misinformation themselves. The thought could be that each agent has their own responsibility to combat conspiracy theories and fake news in order to hopefully prevent their peers from falling down the epistemic rabbit hole of this ignorance and false belief. Just as a libertarian might propose that governmental interference is unnecessary in fixing potholes or building bridges, so too might analogous considerations apply in the no-platforming and/or censorship cases—we can do it ourselves. And, of course, this would also have the helpful upshot of side-stepping the legitimate worries we have about giving powerful institutions such informational control.

I do not find this to be a particularly compelling objection, however. First, we should note that the scenario outlined above is essentially how things *currently* are; governments do not really have much (if any) power to halt the spread of misinformation, fake news, and conspiracy theories (particularly on social media); and corporations, while having almost full autonomy over the running of their platform, apply their rules and community guidelines inconsistently and (often) ineffectively.[25] Moreover, as has been touched on throughout, we are *currently* in an infodemic—if people could effectively combat misinformation themselves, then the contemporary situation presumably should be better—plausibly indicating the ineffectiveness of individualistic strategies. To be clear, individuals should still do their best to combat misinformation in whatever ways they can but it seems to me that the INFODEMIC

25 See, for instance, Mackey and Lee (2022) for discussion of Elon Musk's rule implementation on X (formerly Twitter).

FOLK THEORY gives us reasonable grounds to think that this just will not be enough, and some sort of institutional response may be more effective. But, of course, this then gives rise to the puzzle of epistemic paternalism central in this chapter.

4.4.2 Response 2

Is removing misinformation epistemically paternalistic? Suppose I knew someone was about to lie to you and so before they could speak, I quickly silenced them in some way or another, preventing them from saying their lie. Is this an act of epistemic paternalism from me? Plausibly not: perhaps one could argue that this is not *interference* per se, but instead it is just a morally (or possibly even *epistemically*?) praiseworthy act. The removal of misinformation, then, would be something similar insofar as one's inquiry is merely altered by them no longer encountering testimony or evidence that is, crucially, misleading or false. After all, what is plausibly of philosophical interest about epistemic paternalism is that it is something prima facie epistemically *bad* (hiding true information from an agent, like Goldman's judge) but it results in something epistemically *good* (their epistemic benefit, however that is ultimately cashed out). So, if we buy this objection, then removing falsehoods would not qualify as epistemic paternalism and therefore the censorship and no-platforming I have been discussing throughout is plausibly something entirely different, thus the puzzle dissipates.

Nevertheless, this response does not hold water. For a start, I noted in §4.2 that misinformation is not always false—it can be true information manipulated in such a way to mislead an agent. Therefore, the analogy between stopping a lie being spoken and removing misinformation has already disintegrated and such "true" misinformation *would* qualify as epistemic paternalism by the lights of this objection. Furthermore, we can make this response even stronger. Recall our definition of epistemic paternalism's first requirement, the interference condition: "[The act] interferes with the freedom of inquirers to conduct inquiry in whatever way they see fit." (Ahlstrom-Vij 2013, 61) There is clearly no judgement here on whether the interference is a suppression of information that is true, false, misleading, mendacious, or whatever other quality might be relevant. All that matters, according to this account, is that one's inquiry is interfered with in *any* way.[26] Someone may want to inquire in such a way that includes *all* information on some topic, thus the removal of misinformation would be an interference in their ability to inquire in whatever way they see fit, and (assuming it is done non-consultatively and for the interferee's improvement) thus, still qualify as epistemically paternalistic.

26 For more on this sort of interference and how ubiquitous it may well be, see Medvecky (2020, 82–84). Keep in mind, however, that a mere interference will not necessarily end up being countenanced as *epistemic paternalism* as the other two conditions of non-consultation and improvement would still need to be met.

Here is a counter-response from my opponent: your friend habitually violates Grice's maxims[27] and you know this. Imagine someone asks your friend where the nearest toilet is. You know your friend is going to say "it's just around the corner," but neglect to mention that this toilet is closed. You do something to stop them saying anything in the first place. This does not seem paternalistic (at least, it seems no more paternalistic than when you stop someone lying), yet in this example what they are going to say is entirely true.[28]

This is an interesting case, but overall does not prove worrisome for me here. It is unproblematic because it simply does not qualify as an *epistemically* paternalistic act. Recall the improvement condition on our definition of epistemic paternalism: "[The act] interferes—exclusively or not—for the purpose of making those interfered with epistemically better off." (61) So, a non-consultative interference in one's inquiry that is *not* done for the purpose of the interferee's epistemic improvement would not constitute an epistemically paternalistic act. Applied to this case, the interference at hand is actually making the person epistemically *worse off* as they are not learning some piece of knowledge (the location of a bathroom). Ultimately, this is presumably to their (all things considered) benefit because it prevents them from having false hope and wasting their time due to the bathroom being out of order but, crucially, this benefit is not an *epistemic one*. Arguably this counts as an act of paternalism simpliciter (you are non-consultatively interfering with them to make them prudentially better off) but not its epistemic cousin.

4.4.3 Response 3

Let us briefly take a step back and think about the root of the problem. What the puzzle (and premise IV specifically) highlights is a general worry we would have if the state or other powerful institutions had these epistemically paternalistic powers of no-platforming or censorship—the thought being that we are ultimately motivated by concerns about potential abuses of such powers. Consider the following helpful vignette:

> CORRUPT GOVERNMENT
> During the Tovid-25 pandemic, misinformation, fake news, and conspiracy theories abounded to such a level that it seriously hindered the Schmottish government's response; it adversely affected vaccine uptake, mask wearing, and restriction following. In an attempt to counteract this, legislation was passed which granted the Schmottish government epistemically paternalistic powers of no-platforming and censorship. This was an

27 See Grice (1989), particularly 24–31. Briefly, the four maxims are: quantity—be informative, quality—be truthful, relation—be relevant, and manner—be clear, brief, and orderly.

28 My thanks to an anonymous referee for pressing me on this and giving this case.

> unqualified success and reduced misinformation, fake news, and conspiracy theories to such a level that the government's response was put back on track and they ended up beating the pandemic. Unfortunately, soon after, the Schmottish government became massively corrupt and began using their epistemically paternalistic powers for nefarious purposes—for instance, the silencing of journalists pointing out their corruptness. This meant they could continue in their corrupt ways without consequence.

This sort of case seems the paradigmatic worry behind institutional control of the flow of information. While epistemically paternalistic acts were extremely successful in tackling misinformation, fake news, and conspiracy theories, the same strategies were reused later to conceal corruption, hide *true* information from the public, and importantly, make them epistemically *worse off*. However, there is a problem here. In the first instance when the Schmottish government used no-platforming and censorship policies to combat Tovid-25 misinformation, these were indeed bona fide instances of epistemic paternalism—the Schmottish government non-consultatively interfered with agents' inquiries to make them epistemically better off (viz., to make them not fall prey to Tovid-25 misinformation). In the later instance of the Schmottish government's no-platforming and censorship, however, they were still non-consultatively interfering but, crucially, it was *not* for any agents' epistemic benefit. In fact, it was for the exact opposite reason: to keep them in ignorance of the government's corruption. So, by this failure to meet the improvement condition, the acts are no longer even instances of epistemic paternalism. Thus, if the policies are ever abused, they are—by definition—immediately disqualified from even being epistemically paternalistic. The upshot here is that it is not clear what the puzzle is even about any more, nor if there is even a puzzle at all.

There are a few avenues of response here. First, I think any discussion of epistemic paternalism will be (academically) uninteresting if it is built in that for it to qualify as epistemic paternalism it must be good simpliciter. This also seems to go against general intuitions surrounding epistemic paternalism; for instance, the widespread suspicion of epistemic paternalism would be rather strangely unfounded in the face of this necessary goodness.

For this reason, perhaps an account of epistemic paternalism jettisons the improvement condition and remains only with the interference and non-consultation conditions would more closely capture the cases and the intuitions central to the ideas discussed in this chapter.[29] Sans improvement conditions, fears about the abuses of epistemic paternalism are no longer a category error and are a salient concern. This may strike one as a rather odd account of epistemic paternalism to endorse considering paternalism simpliciter primarily relates to the *improvement* of agents. Nevertheless, from the discussion above, I think there are plausibly prima facie reasons

29 This is an idea McKenna (2020) briefly touches on in footnote 7 of his paper.

that such a two-condition account could be appropriate. It makes sense of our fears of epistemic paternalism outlined in CORRUPT GOVERNMENT and it also plausibly accounts for the general suspicion of epistemic paternalism.

Even so, all things considered, one might reasonably balk at the idea of overhauling the concept of paternalism here. Fortunately, I think there is another, less dramatic route we can take in response to this category error objection. It is important to keep in mind that, throughout this chapter, it has been epistemically paternalistic *policies* that have been the subject our approbation and opprobrium. Although if they overstep there is a sense in which they no longer *strictly* count as epistemic paternalism, they are still *epistemically paternalistic* in nature (and description). Indeed, the very reason for their inception was in the name of epistemic paternalism—that is, the epistemic benefit of interferees. No-platforming and censorship are epistemically paternalistic practices when responding to misinformation and, as potential implemented policies, it is an area worthy of discussion what this might look like when abused even if it would seem to remove the acts from the specific category epistemic paternalism as Ahlstrom-Vij (and myself here) are conceiving of it. The potential problems that could arise from abuses of the epistemically paternalistic policies could not occur if the policies did not exist, and the reason the policies exist is *because of* epistemic paternalism. Analogously, suppose someone has a terrible drinking problem and it eventually led to them injecting alcohol into themselves. It would be asinine at this point to simply shrug and say, "well, they don't have a *drinking* problem any more, they have an *injecting* problem." The issues are clearly connected even though injecting alcohol is not strictly *drinking* alcohol—the injecting issue only exists because of the initial drinking problem. In sum, then, the puzzle is still one of epistemic paternalism irrespective of whether we might think an abuse disqualifies the act from the category of epistemically paternalistic practices. And, most importantly, the puzzle remains of interest.

4.5 CONCLUDING REMARKS

The question of the best or most effective strategy for tackling misinformation is not one that can be easily answered. I proposed that a plausible solution could be that of widespread epistemic paternalism through the practices of no-platforming and censorship. I showed, however, that pursuing this strategy leads to a previously unnoticed puzzle for those who may support the implementation of epistemically paternalistic policies, meaning they ought not be used. I considered some responses and objections to my construction of the puzzle but found each of them wanting, concluding that the puzzle stands. This makes for a troubling conclusion, however. Misinformation, fake news, and conspiracy theories do not appear to be going away any time soon and, if my puzzle is correct, then our best weapon against these wor-

risome issues may prove to be unusable and thus misinformation will continue to be widely propagated among agents in the epistemic environment.

Linking Interlude IV

Throughout this thesis, I have employed many different events, stories, and theories as examples such as the WHO trying to assassinate politicians, climate change denial, Holocaust denial, anti-vaccine sentiment, among many others. Plausibly, what links all these, one might reasonably think, is that they are all, in some sense, *conspiracy theories*. Yet, so far, I have scarcely mentioned the phrase "conspiracy theory", and instead stipulated different categories and definitions (e.g., hedged bullshitting or *controversial false assertions*). This might be somewhat surprising: *Fighting falsity* I so titled my thesis, and, you might think, what better example of such falsity than conspiracy theories? Well, remarkably, the philosophy of conspiracy theories would mostly disagree with this sentiment as it tends to view conspiracy theories and theorising in a more positive light than most (in academia *and* within the laity). Thus, my reticence so far in employing the term. Up until now, that is. This final chapter is a jeremiad of the philosophy of conspiracy theories.

Chapter 5

Disentangling the debate in the philosophy of conspiracy theories: definitions and desiderata

In the philosophy of conspiracy theories, two main views exist: generalism, which holds that conspiracy theories can be negatively epistemically assessed as a class, and particularism, which contends that there is nothing inherently wrong with belief in conspiracy theories and individual conspiracy theories should be assessed on their own merits. The debate between which of these views is to be preferred, however, mostly hinges on differing definitions of "conspiracy theory" meaning each side starts with a radically different idea of the exact concept at stake. In this chapter, I propose a way forward from this impasse. I reconstruct two implicitly endorsed desideratum for any account of conspiracy theories from the extant philosophical literature. I then defend two surprising theses: first, the particularists' favoured definition does poorly on the theoretical desiderata that were supposed to vindicate it; second, the much maligned generalists do much better than has previously been suggested. With these results in mind, I offer up a novel *position to know*-based definition that I suggest charts a path between the Charybdis and Scylla of generalism and particularism.

5.1 Introduction

The philosophical study of conspiracy theories throughout its relatively short lifespan so far has been primarily dominated by two main debates. The first has been a definitional dispute: that is, what exactly is a conspiracy theory, what does a good definition of "conspiracy theory" look like, what do we want from a definition of conspiracy theory, and various interrelated questions. The second has been a battle

between GENERALISM and PARTICULARISM.[1] GENERALISM is the position that conspiracy theories should be assessed as a class (usually as irrational or epistemically faulty in some way) by virtue of their being a conspiracy theory; PARTICULARISM is the position that individual conspiracy theories should be appraised on their own merits (i.e., warrant, evidential basis), and should not be discarded out of hand just for being a conspiracy theory. While these issues might strike a reader unfamiliar with the literature as probably insoluble (like many debates in philosophy), a potentially surprising feature of this story is that both debates have approached near consensus in recent times. The MINIMALIST DEFINITION of conspiracy theory[2] (where a conspiracy theory is simply an explanation of an event which alleges a secret plan enacted by some agents to bring about some end) is now a standard starting point in the philosophy of conspiracy theories, and the particularists have essentially declared victory over the generalists. Indeed, it is almost a rite of passage now to open philosophy of conspiracy theory papers with a declaration of this sort, such as, "It is highly unusual for philosophers to agree about anything. And yet, philosophers of conspiracy theories seem to have achieved this remarkable feat," (Boudry and Napolitano 2023, 22) or:

> A remarkable feature of the first generation of work on conspiracy theory in philosophy has been that—with few exceptions—most philosophers working in the field agreed that there is nothing inherently problematic about belief in conspiracy theories [as theories that allege conspiracies]. (Dentith 2023a, 522)

Or:

> Quite astonishingly, something like a broad consensus has emerged: regarded simply as explanations, conspiracy theories are not intrinsically irrational, and believing in conspiracy explanations is not necessarily unwarranted. (Stokes 2018, 25)

To give just a few examples. A crucial detail to consider, however, is that these are not two discrete debates that coincidentally lulled. In fact, as the Stokes quote tacitly indicates with "regarded simply as explanations," the definitional dispute and the GENERALISM/PARTICULARISM debate are intrinsically linked. Almost every paper which endorses PARTICULARISM also endorses the MINIMALIST DEFINITION (e.g., Pigden (2018), Basham (2018a), and Dentith (2014, 2016, 2018, 2019, 2023a), to give some key players)[3] and every defence of GENERALISM begins with a different (in

1 These terms of art are owed to Buenting and Taylor (2010) and have been widely adopted in the literature.

2 Sometimes also referred to as the NEUTRAL DEFINITION or the SIMPLE DEFINITION. I will use MINIMALIST DEFINITION throughout.

3 There are three exceptions to this rule: Coady (2012), Hagen (2022), and Brooks (2023), who are particularists but have endorsed the MINIMALIST DEFINITION plus a condition of being contra official

some sense pejorative or evaluative) definition of conspiracy theory.[4] Thus, it is no accident that the MINIMALIST DEFINITION becoming orthodoxy was accompanied by the PARTICULARISM turn, nor that non-neutral definitions adding in alternative, pejorative conditions go along with GENERALISM. For the former, it would be rather odd to define conspiracy theories as normatively neutral and then maintain that they are still to be dismissed out of hand as a class; for the latter, it would be incoherent to define conspiracy theories as irrational or epistemically faulty and then contend that they must nonetheless be evaluated individually on their own merits.

So, the debate on whether PARTICULARISM or GENERALISM is correct has essentially hinged on the sort of definition that is started with, and the question of which definition is to be preferred has supervened on whether one's loyalties lie with PARTICULARISM or GENERALISM. Particularists and proponents of the MINIMALIST DEFINITION often accuse generalists of stacking the deck in their favour by presupposing a pejorative definition of conspiracy leaving room for nothing except GENERALISM while generalists and proponents of non-neutral definitions frequently accuse particularists of stacking the deck in *their* favour by presupposing an implausibly broad definition of conspiracy theory leaving room for nothing except PARTICULARISM.[5] The battlelines of the debate being arranged so has had a few upshots, not the least of which is that philosophical argument between the two sides has often proved rather inconsequential as they start from such drastically different grounds about the exact ideas at stake and therefore argue directly past one another.

Here is the plan for this final chapter: in §5.2, I reconstruct two desiderata for any account of conspiracy theories from the extant literature. In the following section, §5.3, I evaluate the MINIMALIST DEFINITION and a definition from Harris on those very desiderata. I show that the former, surprisingly, does poorly, while the latter is generally successful. I nevertheless flag a couple of problems with the latter account

explanations. In practice, however, they do not have any substantive disagreements with any of the PARTICULARISM project. There is another not negligible class of scholars who argue that conspiracy theories are necessarily malevolent (in some way or another) but *epistemically* neutral (Hauswald 2023; Shields 2022, 2023; Räikkä 2018). Despite this, they still fall under the category of MINIMALIST DEFINITION proponents who are particularist as PARTICULARISM relates to epistemic features of conspiracy theories, not ethical ones.

4 There are many examples of this: Cassam (2019a)'s capitalised Conspiracy Theories have a plethora of pejorative conditions; Levy (2007)'s conspiracy theories are contra relevant epistemic authorities; Mandik (2007)'s *proper secrecy* condition; Schaab (2022)'s *heterodoxy* condition; Boudry (2022)'s epistemic black holes; and so on.

5 When I first wrote this in 2022, the connection between definition and general position was merely obliquely referenced by some authors; for instance, "it is fair to say that some beliefs about the likeliness or unlikeliness of conspiracy theories hinge on finessing or questioning such a minimal definition of what counts as a conspiracy." (Dentith 2016, 577) or, "Given the [MINIMALIST DEFINITION], it is understandable that particularists think conspiracy theories are not especially problematic." (Harris 2022, 446–447) In the years since, this connection has been explicitly discussed in the literature by Boudry and Napolitano (2023), writing: "either position can be trivially vindicated by adopting the right definition of "conspiracy theory." (23) Cf. Dentith and Tsapos (2024) for a response to Boudry and Napolitano. In any case, the general point *has* been underappreciated in the literature thus far, even if the tides are beginning to turn, and so it is worth emphasising.

and thus put forward my novel POSITION TO KNOW DEFINITION in §5.4 which I argue is superior on all fronts.

5.2 WHAT DO WE WANT FROM A DEFINITION OF CONSPIRACY THEORIES?

5.2.1 No Trivial (Ir)rationality

In 2014, M Dentith published their monograph *The Philosophy of Conspiracy Theories* (2014). While there are (what are taken to be) key works before this one,[6] I think this book is by far the most influential and seminal piece of work in the philosophy of conspiracy theories. I note this because it is from here that I will mostly reconstruct the desiderata for any account of conspiracy theories. Thus, despite the amount of work in this literature dramatically increasing in size since 2014, it is from this book that most of the literature's axioms have stemmed.[7]

Dentith writes early on: "the central question which underpins the analysis [in the book] that follows is this: 'When, if ever, is it rational to believe a conspiracy theory?'" (5) This is an idea still prevalent in much of the contemporary literature and is an important part of the motivation for the MINIMALIST DEFINITION and PARTICULARISM *and* the case against GENERALISM. As Dentith (and Tsapos) later summarise the general position:

> [I]f we are interested in diagnosing what (if anything) is wrong with belief in conspiracy theories, we cannot start out by assuming that such beliefs are deservedly suspicious. (Dentith and Tsapos 2024, 49)

The idea here is that a key guiding question in the philosophy of conspiracy theories involves asking whether it is (ir)rational to (ever) believe conspiracy theories. Therefore, as it is such a crucial line of inquiry, we do not want to give any old answer to it; we want to give a *good* answer to the question.[8] A bad answer, according to these scholars at least, would be if we *defined* conspiracy theories as irrational to believe (among other conditions) because it would seem to problematically beg the question and give a quick and trivial answer: *it is irrational to believe conspiracy theories because*

6 Beginning with Popper (1966, 94–99) and his brief takedown of what he called the "conspiracy theory of society", followed by Pigden (1995)'s response decades later. Finally, a few years following that, Brian Keeley published "Of conspiracy theories" (1999). These were the only works in the philosophy of conspiracy theories for a number of years, and certainly helped spur philosophical interest in the topic.

7 Cf. Dentith and Tsapos (2024). I will also draw from this paper but it is essentially a retelling of chapters three and four of Dentith's book.

8 This is often referred to as "theoretical fruitfulness" which presumably guides our definitions insofar as a "theoretically fruitless" definition would likely be considered bad. See, e.g., Tsapos (2023) and Duetz (2024). I will also discuss theoretical fruitfulness later in §5.2.3.2.

conspiracy theories are (by definition) irrational to believe. Setting aside any metaphilosophical worries about what exactly makes for a good or bad answer to such a question, at the very least I agree with the antecedent in the quote above that inquiries relating to the (ir)rationality of believing conspiracy theories is an important question in the philosophy of conspiracy theories. If we buy the consequent as well, we might think a desideratum of the following shape is appropriate:

> NO TRIVIAL IRRATIONALITY
> A definition of conspiracy theories ought not trivially entail that belief in conspiracy theories is irrational.[9]

Although not explicitly couched in terms of *desiderata*, this idea that conspiracy theories ought not be considered irrational by definition is commonly endorsed throughout the literature (although sometimes for different reasons than the answering of the rationality question posed above—e.g., the second and third quotes below). Here are some such examples:

> "[Some philosophers' definitions] rule out of court the possibility that belief in a conspiracy theory could ever be rational, which means our analysis is askew from the outset." (Dentith 2014, 28)

> In response [to generalists saying conspiracy theories are epistemically problematic], particularists point out that conspiracies do occur; so, [conspiracy theories] cannot be prima facie unwarranted (by definition) as we know that conspiracy theories about well-known conspiracies are true. (Duetz 2023, 440)

> The primary argument for [particularism over generalism] is that there are an enormous number of conspiracies that we know obtain from the historical record and in our contemporary political lives. It would therefore be an epistemic and political disaster were we to indict such views a priori [i.e., go with a definition that trivially entails irrationality]. (Shields 2023, 467)

Nevertheless, NO EASY IRRATIONALITY (on its own at least) cannot be a bona fide desideratum as it is as problematically one-sided as its apparent target. The thought underpinning *why* it is problematic to define conspiracy theories as "irrational to believe" is that it gives a trivial and altogether too easy (read: *bad*) answer to the question of whether conspiracy theories are ever rational to believe. But notice the problem here is not inherently to do with the *irrationality* aspect, the problem is to do with the

9 Even setting aside this rationality question from Dentith, we might also just think a definition that entails irrationality is bound to be inadequate—presumably there is going to be *some* case where believing a conspiracy theory is not irrational.

trivial answer aspect. Therefore, a definition that trivially entailed that belief in conspiracy theories is *rational* would be problematic by the same question begging lights as the *irrational to believe* definition. So, the desideratum that is really in the backdrop of the above is the following:

> NO TRIVIAL (IR)RATIONALITY
> A definition of conspiracy theories ought not trivially entail an answer about the rationality of belief in conspiracy theories.

This desideratum rules out definitions that trivially entail irrationality *or* rationality.[10] Although the latter has not been explicitly discussed in the literature (likely because no one has put forward an account that says conspiracy theories are rational to believe *by definition*),[11] NO TRIVIAL (IR)RATIONALITY is clearly in the backdrop of all the lines of thought which impugn definitions for trivially entailing irrationality. Therefore, we have our first desideratum. Moreover, it is clearly endorsed (if not explicitly then certainly implicitly) and wielded by many of the particularist, MINIMALIST DEFINITION proponents so they do not have any scope to suggest that they are not beholden to it. As we will see later, NO TRIVIAL (IR)RATIONALITY proves problematic for the MINIMALIST DEFINITION. Before that, however, let us consider another desideratum.

5.2.2 No View Entailment

In §5.1, I described GENERALISM as the view that conspiracy theories can be negatively epistemically assessed as a class while PARTICULARISM contends that this is false and individual conspiracy theories should be appraised on their own epistemic merits. These are widely accepted general glosses of the positions. Notice that this is a discussion of the *theories* themselves and does not make any specific reference to *beliefs*. Confusingly, however, these glosses are not uniform across the literature. In fact, Dentith (again, whom I take to be the leading and most influential philosopher in this field) repeatedly takes the views as also being about *belief* in conspiracy theories:

> Particularists, however, argue that the rationality of belief in conspiracy theories can only be assessed by considering the evidence for and against *individual* conspiracy theories. (Dentith 2016, 582)

> The aforementioned philosophers [key particularists] have argued that there is nothing inherently irrational about belief in such theories when

10 It also would presumably rule out definitions that trivially entailed *arationality* but, so far as I can tell, this is a position no one has discussed in the literature so I will set it aside. The closest is Gadsby (2023) who thinks bad beliefs (in the Levy (2021) sense) stem from arational processes. Nevertheless, while bad beliefs and conspiracy theories can sometimes align, they are certainly not identical.

11 Note, however, that this does not mean a definition could not still *trivially entail* that it is rational to believe conspiracy theories.

> properly understood. Rather, such theories are as good or bad as the evidence which counts for or against them. As such, they have adopted what has come to be called in *conspiracy theory theory* (the academic study of conspiracy theory) 'particularism'. (Dentith 2022, 2)

> [A] *generalist* research programme: generally there is something wrong with belief in conspiracy theories, and this gives us a rationale for treating such theories as *prima facie* irrational. (Dentith 2023b, 406)

> [M]ost philosophers working in the field agreed that there is nothing inherently problematic about belief in conspiracy theories. This consensus has come to be known as 'particularism'. (Dentith 2023a, 522)

> [P]articularists are not so much interested in the definition, but how we should talk about belief in conspiracy theories. (Dentith and Tsapos 2024, 51)

A key point to note here is that features of the theories themselves and the rationality of those who believe them, while often related, quite clearly come apart. For example, assume a definition of conspiracy theories that says, among other things, they are all false. This might make it *less likely* that belief in a given conspiracy theory is rational, but it obviously does not guarantee this. For example, my extremely reliable and trustworthy friend tells me about a conspiracy theory (which we've just said is by definition false, even if myself and my friend are unaware). It would be perfectly rational for me to believe this (absent defeaters, which for the sake of the case, we will say *are* absent). Therefore we have a case where we have a generalist gloss of conspiracy theories (they are assessed as a class as being false), but we have a particularist gloss of *belief* in conspiracy theories (it is at least sometimes rational to believe conspiracy theories due to the certain circumstances in play). We can really distinguish between two different GENERALISMS and PARTICULARISMS, a *theory* version and a *belief* version:[12]

> GENERALISM$_t$
> Conspiracy theories can be negatively epistemically assessed as a class (typically as, e.g., lacking in evidence or some other epistemic flaw).
>
> PARTICULARISM$_t$
> Conspiracy theories should be assessed individually on their own (epistemic) merits.

And:

12 This is also suggested in Harris (2018).

GENERALISM$_b$
Belief in conspiracy theories is inherently epistemically problematic.

PARTICULARISM$_b$
There is nothing inherently epistemically problematic about belief in conspiracy theories.

Which is the "true" GENERALISM and PARTICULARISM and which is the imposter? It is difficult to say. The first quote from Dentith above seems to imply a combination of both PARTICULARISMs outlined here. On the other hand, the thesis put forward in there looks mistaken; for example, imagine conspiracy theory X which objectively has no evidence in support of it. Now suppose I grew up in a cult where X was a central tenet of its teachings. Intuitively, my belief in X is rational despite there being no (non-misleading) evidence for the conspiracy theory.[13] It also seems unlikely that Dentith is imagining a purely subjective conception of evidence either; if so, most conspiracy theories believed in by conspiracy theorists would likely come out as rational as it is often easy to have (bad or misleading) evidence for positions.[14] Accounts of rationality or evidence have never been discussed by the authors in this literature so it is not clear what the intentions are. Keith Harris (one of few avowed generalists) has noted similar confusions and concluded that: "A natural approach is to understand particularism as comprising a constellation of theses." (Harris 2022, 11) He lays out a *weak* and *strong particularism*:

Weak Particularism
The proper assessment of any given conspiracy theory supervenes on the evidence for and against that theory.

Strong Particularism
The proper assessment of any given conspiracy theory supervenes on the evidence for and against that theory *and* there is no evidence against the truth of conspiracy theories as a class. (13)

Notice that his *weak particularism* is essentially in contravention with GENERALISM$_t$ and a recasting of PARTICULARISM$_t$ while his *strong particularism* is close to a combination of PARTICULARISM$_t$ and PARTICULARISM$_b$. I generally agree with this interpretation and think that most particularists hold this combination of PARTICULARISM$_t$ and PARTICULARISM$_b$. In fact, I suspect they tend to hold a position that is something like: PARTICULARISM$_b$ therefore PARTICULARISM$_t$. This does look somewhat plausible if you buy into key particularist theses. For instance, if you take conspiracy theories to

13 This case is similar to Goldman (1988)'s *benighted cogniser*. See Simion, Kelp, and Ghijsen (2016) for some more recent discussion.

14 Conspiracy theories are especially good at having bad or misleading evidence for their position as an often standard part of their evidence set is evidence that *defeats* the justification for competing accounts of the event at hand. See, for example, Keeley (1999) or Napolitano (2021) for discussion of this "self-sealing".

be what the MINIMALIST DEFINITION entails, then it does seem like individual conspiracy theories should be evaluated on their own merits. As I will discuss in more detail shortly in §5.2.3, official accounts of 9/11 or Watergate count as conspiracy theories under this definition thus it would certainly be a misdiagnosis to discard these cases of factual public record out of hand.

Interestingly, on the generalist side, the parallel relation does not seem to hold—GENERALISM$_b$ therefore GENERALISM$_t$ seems more problematic than the particularist version as it could be the case that all those who believe conspiracy theories do so irrationally (as much of social science and psychology holds) whilst (some of) the theories themselves are perfectly legitimate in terms of, e.g., evidential relations or warrant. In any case, while some generalists do seem to hold GENERALISM$_b$ (e.g., Napolitano (2021) and Cassam (2019a)), I think this is mistaken. As my cult case just demonstrated, it is clearly *possible* to have rational belief in a conspiracy theory.[15] Nevertheless, as the trustworthy friend case before that showed, that does not mean we cannot still hold that conspiracy theories can be assessed as a class—i.e., think that GENERALISM$_t$ is true.

The reason why this is important is that, if these belief variants are bona fide instances of PARTICULARISM and GENERALISM (which they certainly seem to be), they clearly connect to our first desideratum: NO TRIVIAL (IR)RATIONALITY. In fact, another desideratum essentially falls out of the first so long as PARTICULARISM$_b$ and GENERALISM$_b$ are legitimate views:

> NO VIEW ENTAILMENT
> A definition of conspiracy theories ought not trivially entail either PARTICULARISM$_b$ or GENERALISM$_b$.

Note that the versions of the two views are the *belief* ones. This is because the theory versions *have* to be entailed by your given definition so they could not feature in a NO VIEW ENTAILMENT-style desideratum. This also plausibly gives us some indication of what the GENERALISM and PARTICULARISM of proper interest are: the belief ones are the only ones that are up for substantive debate irrespective of the account opted for. Thus, again under these "theoretical fruitfulness" considerations, if we want to well answer the question of what view to endorse, we ought not be giving a trivial entailment to either view—our account and our view should be suitably separate. Elsewise, it is simply the definition doing all the work in whether we ought to be generalist or particularist about conspiracy theories and not any argumentation. Again, this general thought has been alluded to, but mainly only as problematic from the

15 It is possible that we could read GENERALISM$_b$ as saying "*typically* problematic", which is more reasonable. Nevertheless, this type of strategy is problematic for the generalist as it means their position can be assimilated into the particularist position; the idea is that as soon as you stop talking about *all* conspiracy theories, the particularist says, "See! You agree that *all* conspiracy theories cannot be assessed as a class therefore particularism is true." See Dentith (2023a) for the implementation of this exact strategy.

generalist side. For instance, as Dentith and Tsapos disapprovingly note: "if we were to work with a [sic] evaluative/pejorative definition, then by fiat generalism seems to be entailed." (Dentith and Tsapos 2024, 49) Therefore, as NO VIEW ENTAILMENT says, if a definition entailing GENERALISM$_b$ ought to be considered problematic then so too should a definition that entails PARTICULARISM$_b$. So, we have our second desideratum that is already implicitly endorsed in the literature *and* independently plausible as a consideration our definitions of conspiracy theory should be held to. With these two negative desiderata, I will argue that, even though (versions of) them have previously been wielded in *support* of the MINIMALIST DEFINITION, it actually fails on both grounds. First, however, it is worth briefly discussing a desideratum that will *not* be showing up.

5.2.3 Extensional Adequacy

In philosophy, our conceptual analysis/definitional projects usually seek *extensional adequacy* as a key desideratum.[16] That is, with conspiracy theories here, we might think the following is an obvious condition on defining conspiracy theories:

> EXTENSIONAL ADEQUACY
> A definition of conspiracy theories ought to match as closely as possible our intuitions and ordinary usage of what counts as a "conspiracy theory".

Perhaps surprisingly, however, co-extension with ordinary language and intuitions has been almost entirely rejected by the particularist, MINIMALIST DEFINITION wing of the literature (which, as noted in §5.1, also makes up the majority *of* the literature) as an unimportant and inessential project for the philosophy of conspiracy theories. Recall that the MINIMALIST DEFINITION says that a conspiracy theory is simply an explanation of some event that alleges a conspiracy—where a conspiracy is just an end desired by some agents who are trying to minimise awareness.[17] This has the upshot, as Harris succinctly puts it, that: "[the MINIMALIST DEFINITION] fails to distinguish between conspiracy theories and non-conspiracy theories in a way that aligns, even approximately, with ordinary usage of the term." (Harris 2022, 3) For instance, surprise birthday parties, the official, received account of 9/11 (where Al Qaeda were responsible), and the Watergate scandal, all come out as *conspiracy theories* on the MINIMALIST DEFINITION. Clearly, this is a huge violation of intuition—and it *is* acknowledged by its proponents (e.g., Dentith (2014, chap. 4)). Various arguments have

16 For instance, the project of analysing knowledge has often hinged on intuitions about Gettier cases and the like. Different projects, of course, have different aims; conceptual engineering/amelioration (e.g., Haslanger (2012) or Manne (2017)) would not aim at extensional adequacy like this as it is doing something quite different. Similarly for stipulative definitions used for theoretical purposes—which is what proponents of the MINIMALIST DEFINITION purport to be doing.

17 See, for instance, Dentith (2016, 577–578). They give three conditions for a conspiracy: the conspirators condition, the secrecy condition, and the goal condition.

been put forward to justify this move which I will now review (and argue they are unconvincing).

5.2.3.1 "That's just a conspiracy theory!"

Despite the MINIMALIST DEFINITION generally being put forward as a *stipulative definition* for theoretical purposes (e.g., Dentith (2014) with the theoretical purpose being fruitful analysis of the (ir)rationality of conspiracy theories), the most common argument in favour of jettisoning EXTENSIONAL ADEQUACY actually has a *conceptual engineering* flavour. The general idea is that it is bad when certain individuals (i.e., bad actors, actual conspirators, politicians) disregard sound questions or inquiries with the response of "that's just a conspiracy theory!" Thus, we should advocate for and adopt a neutral definition (a la the MINIMALIST DEFINITION) to go some way in preventing this negative consequence as we hopefully make it less likely that people view conspiracy theories as "bad" (be that epistemically, morally, etc.) and so less likely to wield it in such a manner. Crucially, this may ameliorate worries about this negative use of "conspiracy theory" which plausibly leads to conspirators getting away with *real conspiracies*. As Dentith writes:

> If we preserve the notion that the terms 'conspiracy theories' and 'conspiracy theorists' are pejoratives, then this might shield conspirators from the accusation that they are conspiring. (113)

So, because lay intuitions are that conspiracy theories are bad in some way or another,[18] then the pursuit of EXTENSIONAL ADEQUACY would only solidify this negative conception which, purportedly, "invests the lies, evasions and self-deceptions of torturers and warmongers with a spurious air of methodological sophistication." (Pigden 2006, 33) Thus, we should jettison EXTENSIONAL ADEQUACY and any definitions of conspiracy theory that include some epistemic fault, and instead, according to most, opt for the MINIMALIST DEFINITION. This argument originates with Pigden (2006; 2007; 2016) where he regularly employs the example of Tony Blair dismissing (appropriate) questions into the Iraq war by calling them conspiracy theories.[19] Since then, it shows up in the many particularist leaning papers in the philosophy of conspiracy theories as a key reason for the adoption of the MINIMALIST DEFINITION and so is a vital argument for the endorsing of the MINIMALIST DEFINITION *and* the rejection of EXTENSIONAL ADEQUACY.[20]

This argument is nonetheless rather unpersuasive for a couple of reasons. The first is straightforward; as Scott Hill puts it:

18 See Napolitano and Reuter (2023) for empirical data supporting this.

19 Although see Hill (2024) for convincing reason to think even this case is not a particularly good example of the phenomenon Pigden is trying to highlight.

20 See, for example, Basham and Dentith (2016) and Hagen (2024, 2025) for employment of this argument. See also Hauswald (2023) for an explicit discussion of what he calls the *dismissive conversational exercitive* of "That's just a conspiracy theory!"

> If 'conspiracy theory' were a neutral term and instead [X reasonable conspiracy theory] was called 'preposterous' by the media, I don't think that would have changed the degree to which [X reasonable conspiracy theory] was unfairly dismissed. (Hill 2024, 50)

The general idea is the following: suppose that the MINIMALIST DEFINITION (or any other suitably non-pejorative definition) were to be adopted by everyone such that the phrase "That's just a conspiracy theory!" or similar dismissals were no longer felicitous and so fell out of use. It is not obvious that this would do much at all to help prevent legitimate inquiries from being disregarded by the same bad actors. It is not as though once "that's just a conspiracy theory" is removed from a bad (or otherwise) actor's toolkit they no longer have *any way* to disparage some theory or line of inquiry—as Hill says above they could call it "preposterous" presumably to similar effect. In fact, as I see it, there are plausibly *more* effective dismissals that do *not* invoke conspiracy theories or theorists; for instance, saying that something "has already been debunked" or that "no serious people think that" strike me as rather good ways of signalling that some idea is nonsense that may be similarly inappropriate insofar as the idea may not have been debunked and may be perfectly legitimate for serious people to believe.

For Pigden's argument to work, it needs to be the case that there is something distinctly effective about dismissing an idea with allegation of conspiracy theory compared to any other dismissal we can come up with. There is no evidence for this and in fact some evidence of the *opposite*: empirical data from Michael Wood suggests that calling something a "conspiracy theory" compared with merely an "idea" does nothing to prejudice people against it (Wood 2015). So, the argument that we ought to induce change in the concept 'conspiracy theory' away from pejoratives due to dangerous dismissals of legitimate inquiries fails on two key grounds: should the conceptual change project be successful, conspirers would likely still be able to inappropriately dismiss proper questions with different terms; and empirical data suggests the dismissal may not even be effective at causing the problem supposed to vindicate such advocation for conceptual change.

The above alone is likely fatal to this conceptual engineering argument. Nevertheless, let us set aside this response, and grant Pigden's and Hagen's (among others) assumption that "conspiracy theory" allegations are some uniquely effective way of getting away with conspiring/dismissing legitimate inquiries *and* disregard the empirical data from Wood. Even with these (extremely) charitable concessions, I *still* maintain the argument does not constitute good reason to try to change the meaning of the term and adopt the MINIMALIST DEFINITION. Zooming out, the general line of argument underpinning this conceptual engineering proposal is (presumably) that misuse or abuse of a term should entail that we (try to) change the meaning of that term to prevent this. The problem is that this reasoning clearly does not follow; it is

not *at all* obvious why a (deliberate or not) misapplication of a term gives normative grounds to endorse changing the meaning of the term.

For example, in right-wing politics (especially with American republicans), it is common to label certain policies such as Medicare for All as "communism" (Müller 2024) or say that the Nazis were "socialists" (Granieri 2020). Both of these are misuses of the term—Medicare for All is not communism and the Nazis were not socialist—that are employed to attempt to illegitimately get their audiences to disregard the policies (universal healthcare) or movements (socialism) in play. Thus, these uses are analogous with the above complaint about conspiracy theories. Let us also assume (as we are now with the "conspiracy theory" dismissal) that these are genuinely effective in prejudicing people against that which the label is being misapplied. If we buy Pigden's argument for changing the meaning of "conspiracy theory", then that would imply that we ought to change the meaning of "communism" and "socialist" too in order to prevent similar abuses. This is, of course, untenable and nonsensical and no one would ever advocate for such a move. The appropriate response is rather clear: we criticise and denounce those who infelicitously employ the terms like this and make clear that these are unscrupulous and deliberate misuses of them to further some agenda. Nowhere does making an attempt to change the meaning of the at-issue terms fit in. So, even if we allow that allegations of conspiracy do have such deleterious effects, the argument underpinning the conceptual engineering proposal does not stand to reason when generalised.

5.2.3.2 Theoretical fruitfulness

A second argument commonly fielded for the rejection of EXTENSIONAL ADEQUACY[21] is an appeal to "theoretical fruitfulness" (Dentith 2014; Duetz 2023, 2024). The idea is that it is not "theoretically fruitful" to pursue EXTENSIONAL ADEQUACY; rejecting intuition and ordinary language considerations (usually by making the MINIMALIST DEFINITION central) is the way to do theoretically fruitful work in the philosophy of conspiracy theories. A natural question at this point might be: what do we mean by "theoretical fruitfulness"? For Dentith at least, this question has an answer familiar from §5.2.1: giving a good answer to the question of whether it is ever rational to believe conspiracy theories. The thought was presumably something like the following: if we try to meet EXTENSIONAL ADEQUACY in our definition of conspiracy theories then we necessarily end up with an account that says (among other things) conspiracy theories are irrational (to believe). Therefore, as previously discussed, we do not get a good answer to the rationality question (it is answered by default) and thus our definition is not theoretically fruitful. So, to prevent this result, we ought not pursue

21 And generally the adoption of the MINIMALIST DEFINITION. While one does not necessarily need to endorse the MINIMALIST DEFINITION after rejecting EXTENSIONAL ADEQUACY, these generally go hand-in-hand in the literature.

EXTENSIONAL ADEQUACY and endorse a "neutral" starting point: "the [MINIMALIST DEFINITION] is the most theoretically fruitful option to work with if we are concerned about belief in conspiracy theories." (Dentith and Tsapos 2024, 49)

This is unconvincing, however. At best, it is simply outdated (the brunt of this work is from 2014 and the 2024 work draws heavily on *The Philosophy of Conspiracy Theories*), at worst it is an obvious strawman. We are essentially presented with a false dichotomy: either (i) we chase EXTENSIONAL ADEQUACY and so adopt an account that defines conspiracy theories as irrational and is thus theoretically fruitless, or (ii) we reject EXTENSIONAL ADEQUACY (for the reasons of (i)) and adopt the MINIMALIST DEFINITION which is theoretically fruitful. However, the jump from the pursuit of EXTENSIONAL ADEQUACY to an account that says conspiracy theories are irrational by definition is demonstrably false. For example, Harris (2022, 2023)'s CONTRA EPISTEMIC AUTHORITIES DEFINITION of conspiracy theories does fairly well on the EXTENSIONAL ADEQUACY while being non-pejorative and not entailing that conspiracy theories are irrational. My favoured *position to know* account which I will later give does even better on capturing intuitions and ordinary language and similarly does not give a trivial answer to the rationality question. So, this line of "theoretical fruitfulness" from Dentith is wholly unpersuasive.

Nevertheless, Dentith is not the only one to make reference to theoretical fruitfulness in discussing what account of conspiracy theories to favour. Many others do so (Tsapos 2023; Napolitano 2021; Napolitano and Reuter 2023) but tend to leave what exactly it means unspecified at the level of intuition.[22] Duetz is the exception,[23] writing:

> Such goals, for example, might concern; acquiring a better understanding of conspiracy theories, understanding the people who believe them, analysing their logical structure, examining the possibility of finding defeaters, and proposing solutions for conspiracy theory-induced polarization. (Duetz 2023, 448)

She concludes, as Dentith did, that the MINIMALIST DEFINITION is best placed to achieve these goals, while non-MINIMALIST DEFINITIONs will fail. She does not provide any argument for this, however, aside from the final goal in the quote which was the focus of her paper. We can allow that she is right about that one[24] but the rest

22 This is particularly unhelpful as many (myself included) likely take EXTENSIONAL ADEQUACY to be a way for a definition *to be* theoretically fruitful.

23 To be clear, she is specifically discussing the shortcomings of pejorative accounts of conspiracy theories with reference to the "goals central to conspiracy theory theory". This is broadly synonymous with the "theoretical fruitfulness" I have been discussing here.

24 Although I am sceptical. Her point is that a negative conception of conspiracy theories will only further polarize those who believe them thus we should adopt a neutral definition (i.e., the MINIMALIST DEFINITION). I do not find this a convincing argument; it is akin to saying that we should not have a negative conception of racism as it will only further negatively polarize racists. This is not only implausible but also abhorrent.

clearly do not rely on the MINIMALIST DEFINITION (in fact, I will later argue that the MINIMALIST DEFINITION does especially bad at some of these). For instance, even if we go with the much maligned *conspiracy-theories-as-irrational account*, it is not at all obvious why this is at odds with understanding conspiracy theories or the people who believe them (we could perhaps look at why or how it is that people believe them while they are irrational to believe), or analysing their logical structure (while I'm not exactly sure what this means, one guess is that alternative accounts are simply different ways of analysing the logical structure of conspiracy theories), or finding defeaters (again, why would this be limited to the MINIMALIST DEFINITION?). So, arguing that jettisoning EXTENSIONAL ADEQUACY is vindicated by the "theoretical fruitfulness" garnered by adopting the MINIMALIST DEFINITION is at the very least not sufficiently shown. None of the supposed benefits attained are incompatible with pejorative definitions except for Dentith's rationality question but no contemporary generalist philosopher holds a *conspiracy-theories-as-irrational* account anyway (except possibly Cassam (2019a)).

5.2.4 Taking stock

In sum, we have seen that a number of particularists give a variety of arguments to justify their move away from EXTENSIONAL ADEQUACY which they in turn take to vindicate adopting the MINIMALIST DEFINITION. I have shown that none of these arguments are persuasive. So what now? We might think that this justifies the move to take up EXTENSIONAL ADEQUACY as a bona fide desideratum on definitions of conspiracy theories. I am tempted by this but nevertheless remain reticent. The worry is that as soon as you take seriously EXTENSIONAL ADEQUACY, the particularist wing of the literature just throws their hands up and says you are stacking the deck against them as their account manifestly fails on that front. Therefore, in order to entirely meet the particularist and MINIMALIST DEFINITION proponent on their own grounds, I will not treat EXTENSIONAL ADEQUACY as a proper desideratum and remain merely with NO TRIVIAL (IR)RATIONALITY and NO VIEW ENTAILMENT—which, as I already discussed, *are* endorsed by them (if tacitly or even unknowingly). Nevertheless, I take it that, all else equal, an account is better if it meets EXTENSIONAL ADEQUACY than if it does not—even the MINIMALIST DEFINITION is not *wholly* removed from ordinary language. Thus, we can maintain EXTENSIONAL ADEQUACY in the background as a sort of tie-breaker if accounts meet the two genuine desiderata adequately.

5.3 EVALUATING EXTANT ACCOUNTS

Now that we have our desiderata, we can evaluate two extant accounts on them. Specifically, I will first look at the MINIMALIST DEFINITION. As I mentioned throughout the previous section, (versions of) both desiderata were put forward *in support* of

the MINIMALIST DEFINITION, so the conclusion I will shortly defend is a surprising one: it fails on both grounds. I will then discuss the CONTRA EPISTEMIC AUTHORITIES DEFINITION from Harris (2022, 2023) and show that, also notably, it does perfectly well on both.

5.3.1 The Minimalist Definition

5.3.1.1 No Trivial (Ir)rationality

The MINIMALIST DEFINITION trivially entails the rationality of belief in conspiracy theories. Admittedly, it does not do it *by definition* or *analytically* like the *conspiracy-theories-as-irrational* patsy that has been discussed sporadically throughout did with *ir*rationality. Nevertheless, it is obvious that, upon adoption of the MINIMALIST DEFINITION, belief in conspiracy theories is trivially rational. It is not even necessarily because established facts of history come out as conspiracy theories on the account—recall, the official story of 9/11 where Al Qaeda were the perpetrators, the Watergate Scandal, and so on, all are countenanced as conspiracy theories on the MINIMALIST DEFINITION. Insofar as is rational to believe these stories (it is) we might think this makes the rationality of conspiracy theories rather easy. It does, but if this were all we had to propose the *trivial entailment* of the rationality of conspiracy theories, we might argue that this is not quite robust enough; there is a somewhat contingent flavour here inasmuch as the fact that it is (now) perfectly rational to believe these official stories does not make it *trivial* that belief in conspiracy theories is rational, it just makes it quite plausible.

Nevertheless, this is not all we have. In fact, despite the clear violation of intuition in denoting whatever historical fact—say, D-Day planning—as a conspiracy theory, the definitional weakness of the MINIMALIST DEFINITION runs far deeper than just these sorts of examples. Consider Apple in the early stages of bringing out a new product or a university's Philosophy Department holding a staff discussion of essay questions for the coming exam season. Recall the conditions to be met are simply a group of agents with some end in mind who are trying to minimise awareness. Apple are a (large) group of agents with the end of bringing out a new product who do not want everyone to know about it yet. The Philosophy Department is a group of agents with the end of preparing exams who do not want their students to know the questions in advance.[25] If you believe that any of this has ever happened, is happening, or ever will happen, you believe in a story that alleges a conspiracy and thus, according to the MINIMALIST DEFINITION, believe a conspiracy theory. If you believe that at any point in history there have ever been a group of people who wanted to keep somewhat quiet from some people what they were doing, then you are a conspiracy

25 It is transparently absurd to consider these cases conspiracy theories as well, but that is not the issue at hand here.

theorist. This, more than anything, is trivial.

Recall that the reason we ended up with this NO TRIVIAL (IR)RATIONALITY desideratum was because of Dentith's rationality question: they wanted to give a good answer to their key research question of "is it ever rational to believe conspiracy theories?" The (to my mind, plausible) thought was that a definition that simply entailed irrationality was problematically begging the question. Somehow, however, it has been completely missed that their favoured definition is doing just as trivial work in the other direction.[26] Even more strangely, this triviality should be obvious, particularly when scholars like Pigden write: "we can conclude that every politically and historically literate person is a conspiracy theorist." (Pigden 2016, 130)[27] Pigden is right about this of course—with the MINIMALIST DEFINITION suitably in place—so how could we possibly think this is giving us a good answer to the question of whether it is ever rational to believe conspiracy theories? In practice, the answer given by the MINIMALIST DEFINITION is just as trivial as *conspiracy-theories-as-irrational*'s. So, the MINIMALIST DEFINITION fails on our first desideratum, NO TRIVIAL (IR)RATIONALITY, for the exact same (albeit inverted) reasons that Dentith and compatriots discounted other definitions of conspiracy theories.

5.3.1.2 No View Entailment

PARTICULARISM$_b$ says that *there is nothing inherently epistemically problematic about belief in conspiracy theories.* As we have seen, the MINIMALIST DEFINITION trivially entails the rationality of belief in conspiracy theories. Therefore, presumably, it also trivially entails PARTICULARISM$_b$ and thus fails NO VIEW ENTAILMENT.

One might find this a bit quick, however. Proponents of the MINIMALIST DEFINITION do not view it as entailing PARTICULARISM:

> Indeed, people can keep to a general skepticism of conspiracy theories while operating with a perfectly neutral and non-pejorative take on what counts as a 'conspiracy theory' and who counts as a 'conspiracy theorist'. (Dentith 2014, 173)

Here, Dentith is claiming that it is entirely possible for one to be a generalist (following GENERALISM$_b$, that is) while holding the MINIMALIST DEFINITION as their preferred account. Thus, according to them at least, the MINIMALIST DEFINITION manifestly does not entail either of the two views. Indeed, as mentioned in the above

26 My conjecture here is that, for whatever reason, these scholars (Dentith, Pigden, Basham) only view it problematic or a bad answer if you trivially entail *irrationality*, and see no issue with a *rationality* entailment. That is, they hold NO TRIVIAL IRRATIONALITY: *A definition of conspiracy theories ought not trivially entail that belief in conspiracy theories is irrational.* and do not seem to realise that you cannot hold this without holding a *rationality* entailment equally problematic.

27 Of course, this does not actually go far enough; neither political nor historical literacy is really required to rationally believe conspiracy theories on the MINIMALIST DEFINITION—I suspect even a cursory knowledge of the trashiest gossip magazine would be sufficient.

quote, most view the MINIMALIST DEFINITION as a neutral/non-prejudicial starting point in the philosophy of conspiracy theories.[28] Moreover, if you find my argument in the previous section unconvincing then this failure of NO VIEW ENTAILMENT will not go through. In general, we might want there to be separate arguments that the MINIMALIST DEFINITION fails to meet NO VIEW ENTAILMENT anyway. I will here give two; the first will be of a similar cast to the previous section; the second focuses on contextual evidence from various authors.

Philosophers in this literature tend to assume that GENERALISM and PARTICULARISM fill the conceptual space—viz., if you are not a particularist, you are a generalist, and vice versa.[29] Therefore, if an account is *incompatible* with either view, it would presumably trivially entail the other through a simple disjunctive syllogism. The MINIMALIST DEFINITION is clearly incompatible with the theory and belief versions of GENERALISM so even if PARTICULARISM proper is a combination of *both* the belief and theory versions (as Harris says, and I am also inclined to think), it looks like the MINIMALIST DEFINITION trivially entails PARTICULARISM and thus fails NO VIEW ENTAILMENT.

Perhaps a reasonable complaint here is that I have essentially just repeated the earlier objection about rationality. Fortunately, there is an additional convincing line that demonstrates how the MINIMALIST DEFINITION clearly entails particularism, one that cannot be disputed by MINIMALIST DEFINITION proponents: their own words.

A commonplace occurrence in the philosophy of conspiracy theories is that different scholars who all endorse the MINIMALIST DEFINITION at times treat the definition and PARTICULARISM as essentially one and the same or describe definitions *as being* particularist. Here are some such examples: "Any particularist definition will do here," (Brooks 2023, 3281) "Minimalist/particularist conceptions," (Duetz 2023, 449) "Dentith's minimalist notion of particularism." (Stamatiadis-Bréhier in Dentith and Tsapos 2024, 55) This looks rather like endorsing the MINIMALIST DEFINITION carries with it an endorsement of PARTICULARISM, or even that they are synonymous as Duetz seems to imply. If any of these intepretations are true then it certainly looks like the MINIMALIST DEFINITION trivially entails PARTICULARISM and so it fails our second desideratum. Nevertheless, one might respond that this is all at best circumstantial evidence from plausibly loose talk or slipshod phrasing and I ought not place the weight of my argument on it. Moreover, none of those cited are Dentith who has really been my main target *and* is whom I take to be the key philosopher in this literature. So, perhaps if Dentith made the same implications, *then* this would be a convincing point. As it happens, not only does Dentith make the same implications,

28 I will take a closer look at this "neutral starting point" assumption in the following section.

29 Patrick Stokes (2018) purports to advocate for a "reluctant particularism" or a "defeasible generalism" but these positions inevitably collapse into PARTICULARISM simpliciter because they do not apply to *all* conspiracy theories. *Mutatis mutandis*, Stamatiadis-Bréhier (2023, 2024)'s "local generalism". See also note 14.

they do so in a way that makes it essentially impossible to argue that the MINIMALIST DEFINITION does not fail NO VIEW ENTAILMENT.

Boudry and Napolitano (2023) discusses PARTICULARISM and GENERALISM and, similar to how I put it in §5.1, notes that those on either side frequently argue past each other. Thus, they advocate for getting rid of the positions and terminology, saying:

> Ultimately, we believe that the generalism vs. particularism divide should be abandoned in favour of alternative conceptual maps, in order to foster better, more constructive, philosophical disagreements. (23)

Dentith and Tsapos (2024) disagree, and to argue why we should not abandon this divide, they reference twenty separate pieces of work (52–53) that, they claim, employ or are framed under PARTICULARISM thus proving it is a "a successful epistemic project." (52) What is important for our purposes here is that six of these papers objectively have almost *nothing* to do with PARTICULARISM![30] The first example, Basham (2018b) does not mention the phrase particularist or particularism once. The second, Brooks (2023), has a single mention of particularism in a footnote at the beginning of the paper. For the third, Dentith and Tsapos say that, "Will Mittendorf *uses particularism* to illustrate the roles of epistemic norms in judging the legitimacy of our democratic institutions." (Dentith and Tsapos 2024, 53) [my emphasis] Yet in Mittendorf (2023), he only mentions particularism four times, mostly tangentially, with the most clear mention him saying, "even if the particularists are right... [here is a problem]." (483) It is not clear at all how this is him "using particularism". Stamatiadis-Bréhier (2023) is also cited as "using particularism" when he actually argues against it, instead proposing his "local generalism". Tsapos (2024a) only mentions particularism in the abstract and the introduction and even the aim of the paper takes itself to be raising a problem *for* particularists, while Tsapos (2024b) again has literally no mention of particularism.[31]

So, what explains this littany of errors? Are Dentith and Tsapos lying to bolster the particularist project? Have they not read the work but are pretending to? Have they forgotten what they themselves wrote? I think these explanations are completely implausible. Nevertheless, there *is* one sensible explanation. All the mentioned papers above endorse the MINIMALIST DEFINITION at the outset before discussing issues *unrelated* to PARTICULARISM, GENERALISM, or definitions of conspiracy theory. Therefore, the reason why they feature here in an apparent discussion of the successes of PARTICULARISM is so: if you endorse the MINIMALIST DEFINITION, you (trivially) entail PARTICULARISM. From this textual analysis, we can see that the proponents of the MINIMALIST DEFINITION themselves clearly believe it entails PARTICULARISM and so it must fail the NO VIEW ENTAILMENT desideratum.

30 Possibly more—the exact position endorsed in some of the papers is left unclear.
31 We can only assume that Dentith wrote this section of Dentith and Tsapos (2024).

In sum, then, we can see that, surprisingly, the MINIMALIST DEFINITION clearly fails both desiderata—versions of which were initially endorsed *by* its proponents in order to advocate for it. Thus, I think this proves rather damning for the MINIMALIST DEFINITION. There is one last response I think defenders of the MINIMALIST DEFINITION could field; yes, the MINIMALIST DEFINITION does ultimately trivially entail both the rationality of conspiracy theories *and* PARTICULARISM. Nevertheless, all this proves is that both of these positions are correct, and the desiderata ought to only apply to *ir*rationality and GENERALISM. After all, they say, the MINIMALIST DEFINITION is a non-pejorative, non-evaluative definition; if we start from a position of *neutrality* and end up entailing some views, well it must just mean those views are correct. I will now reject this assumption.

5.3.1.3 A "neutral" starting point?

A point often indirectly touted in favour of the MINIMALIST DEFINITION is its *neutrality*; this also dovetails with my earlier discussions of those scholars that felt it was problematic if accounts entailed irrationality or GENERALISM and so they lauded their definition for not falling prey to such negative evaluations. Here are some such examples:

> Starting from a descriptive, rather than evaluative, conception of 'conspiracy theory' implies a neutral starting point with respect to the rationality of believing such theories, and, hence, does not contribute to conspiracy theory-induced polarization as strongly as does a non-neutral conceptual basis. (Duetz 2024, 2107)

> [B]y remaining neutral about the nature of the appropriate epistemic evaluation in conceptualizing 'conspiracy theory', Minimalist accounts do not limit themselves to focusing on only one account of problematic conspiracy beliefs. (2113)

> . . . by emphasizing the importance of a neutral and minimal starting point for inquiries into conspiracy theories, conspiracy belief, conspiracy mindsets, and so on. (2114)

> [A] perfectly neutral construal of both what counts as a 'conspiracy theory' (any explanation of an event that cites a conspiracy as a salient cause) and a 'conspiracy theorist' (anyone who believes some conspiracy theory). (Dentith 2014, 122)

> I will take a neutral definition of the term to properly investigate what expertise on conspiracy theories really entails. (Tsapos 2024b, 4)

The thought, I assume, is something like: to get true answers in the philosophy of conspiracy theories, we must start from a baseline that does not include any biases

or evaluations—that is, an account that is completely neutral. The issue with other accounts is that they build in evaluative or pejorative features which then prejudice the case on whatever it is we are discussing. The MINIMALIST DEFINITION cannot prejudice anything because it is *neutral*.

I do not find this at all plausible. I am reminded of how centrists often equate their political position with "neutrality" as though that somehow makes it better or more sensible than the left or right, even though centrism is just another political ideology. Here is perhaps a clearer analogy. Suppose I wanted to research the following line of inquiry: when, if ever, is it legally okay to commit murder?[32] I want a good answer to this question, not a trivial entailment from a definition, so I decide to shift from the evaluative, pejorative definition—that says murder is unlawful killing—to a "neutral" definition; say, murder is just any act of killing. From here, I now go about answering my central question, and discover that indeed it is legally kosher to kill—for example, in a war—therefore, murder is (sometimes) legally okay. This answer is not up for opprobrium, I contend, because I started from a completely neutral, non-evaluative, non-pejorative baseline! Therefore, any conclusions drawn are far more legitimate than any definition relying on non-neutral conditions. I take this to be a fairly plausible reductio.

To be clear, I am not arguing here that conspiracy theories are like murder insofar as a clearly negative evaluation ought to be baked in. Rather, I am arguing the weaker point: "neutrality" does not necessarily mean *some evaluation* is not still coded in. Because the MINIMALIST DEFINITION is so broad and easily applicable, it encodes lots of evaluations: like surprise birthday parties being conspiracy theories. Therefore, conspiracy theories can be good and nice. This is an *evaluation*, despite the claim that the definition is "neutral". So, I do not buy this claim that the MINIMALIST DEFINITION is somehow "neutral" and this makes it better than other accounts.

5.3.1.4 Summing up

I have argued that the MINIMALIST DEFINITION fails on both desideratum that were supposed to vindicate it. First, it straightforwardly entails that belief in conspiracy theories is rational, failing NO TRIVIAL (IR)RATIONALITY. It also fails NO VIEW ENTAILMENT which I evidenced using its proponents' own words. Finally, I called into question the widespread assumption of the MINIMALIST DEFINITION's neutrality. In sum, then, it is unsuccessful on the very grounds used to reject a variety of pejorative definitions and so ought to be rejected as a satisfactory account of conspiracy theories.

5.3.2 Contra Epistemic Authorities Definition

Keith Harris (2022, 2023) defines conspiracy theories as follows:

32 This is supposed to be analogous to Dentith's rationality question: when, if ever, is it rational to believe conspiracy theories?

CONTRA EPISTEMIC AUTHORITIES DEFINITION:
A theory, t, is a conspiracy theory if and only if:

1. t is an explanation of events that alleges a conspiracy, and
2. t is in conflict with the claims of relevant epistemic authorities.

The first condition of this account is simply the MINIMALIST DEFINITION (meaning it a necessary but not sufficient condition on accounts of conspiracy theories), while the second is a new, contra epistemic authorities condition.[33] There are two key points in need of explanation here: i) what a "relevant epistemic authority" is, and ii) what it is to "conflict with claims" from them. Let us start with the first.

Harris takes "epistemic authority" to be a matter of "credentials, positions, and the like," (Harris 2022, 7) as opposed to a definition that cashes it out in terms of epistemic superiority as is more common in epistemology.[34] Harris' reason for this is that an epistemic superiority (or reliability) cash out is problematically pejorative insofar as it would make conspiracy theories (and belief therein) *inherently* epistemically flawed. As we concurred in this chapter, such an assumption would be mistaken: it appears at least possible to sometimes have a rational belief in some conspiracy theory. Therefore, to avoid this worry, Harris understands "epistemic authority" in a more *social* manner: "the reliability of epistemic authorities will be contingent upon whether credentials and positions are reserved for those who are epistemically reliable." (7–8) So, in what he calls "well-functioning systems", the reliability and the given credential will adequately match up, insofar as only those who are reliable when it comes to answering questions of mathematics will be given the credential of "mathematician" (e.g., a PhD in mathematics). *Mutatis mutandis* whatever position or credential in play.

On the other hand, in "poorly functioning systems", there may be a disconnect between credentials allocated and given reliability. Harris references nepotism or bribery as potential factors in the credential given to the expert in a poorly functioning system. An example of this might be a previous chief executive of a wrestling entertainment company being made head of education in government—the idea being that such a person is not going to be reliable in their education-centred question-answering like the mathematician was, yet they still have the *credential* of being an epistemic authority by virtue of their position. This allows for some give in the definition such that, in a society with such poorly functioning systems, it might be wholly rational to believe some theory that alleges a conspiracy and conflicts with the relevant epistemic authorities—because such authorities are not necessarily reliable.

33 I say "new", but note that the actual progenitor of this *contra epistemic authorities account is* Levy (2007). See also Levy (2021) for similar ideas although discussing his "bad beliefs" as opposed to conspiracy theories specifically.

34 For the classic work on this, see Zagzebski (2012). Cf. Goldman (1999) and the reliability central there.

Turning now to ii), the *conflict* in play in the definition, Harris leaves it mainly at the intuitive level. I do not think this is particularly problematic, however. Suffice to say the conflict can be both explicit or implicit. A conspiracy theory saying that Covid-19 vaccines are deadly *explicitly* conflicts with the vast numbers of claims, studies, reports from epistemic authorities (doctors, scientists, etc.) worldwide. Nevertheless, some nascent or inconsequential conspiracy theory, might not have any explicit refutation from any epistemic authorities—suppose a conspiracy that says car tyres are shaped such to psychologically trick people into buying more doughnuts. Nevertheless, the thought is that this would still *implicitly* conflict with the claims of relevant experts; viz., engineers who say that tyres are shaped thus for driving reasons. Harris does note that, "questions may remain… concerning what degree of conflict… is required." (Harris 2022, 9). In practice, however, I do not view this as a decisive problem for the account.

With the particulars of the CONTRA EPISTEMIC AUTHORITIES DEFINITION now outlined, we can test the account on our desiderata.

5.3.2.1 No Trivial (Ir)rationality

First, it meets NO TRIVIAL (IR)RATIONALITY. It does not trivially entail anything about the rationality of conspiracy theories. It does not trivially entail that belief in conspiracy theories is *rational*—if anything, it telegraphs a move towards a general suspicion of conspiracy theories, granted we are in a somewhat well-functioning system and so have reliable epistemic authorities. We might worry then that it trivially entails *irrationality*. This is mistaken, however. It might make it more *probable* that beliefs in conspiracy theories will not be rational, but it does nothing to *trivially entail* this. Recall, Harris is careful to note that his socially demarcated epistemic authorities are not infallible nor necessarily countenanced in terms of reliability—only contingently so—thus avoiding this collapse into trivial irrationality when conflicting with their claims.

5.3.2.2 No View Entailment

In light of the above then, CONTRA EPISTEMIC AUTHORITIES DEFINITION also meets NO VIEW ENTAILMENT insofar as it trivially entails neither GENERALISM$_b$ nor PARTICULARISM$_b$. It does not say that belief in conspiracy theories is inherently problematic as it allows for the possibility of reasonable belief in conspiracy theories. Nor does it trivially entail, as the MINIMALIST DEFINITION did, that there is nothing inherently problematic about belief in conspiracy theories as it leans towards a defeasible suspicion of conspiracy theories. So, rather straightforwardly, the CONTRA EPISTEMIC AUTHORITIES DEFINITION meets the second desideratum as well. As it will prove important for later comparison with my novel account, let us now also take a look at

EXTENSIONAL ADEQUACY.

5.3.2.3 Extensional Adequacy

The CONTRA EPISTEMIC AUTHORITIES DEFINITION does far better than the MINIMALIST DEFINITION when it comes to EXTENSIONAL ADEQUACY. Recall the MINIMALIST DEFINITION diagnosed the received account of 9/11, the Watergate scandal, and so on, *as* conspiracy theories, which is universally agreed to be a huge violation of intuition. On the CONTRA EPISTEMIC AUTHORITIES DEFINITION, this is not the case. Neither of the received account of 9/11 nor the Watergate scandal are in conflict with the relevant epistemic authorities, thus do not meet the second condition, and thus are (correctly) not countenanced as conspiracy theories. So, one significant upshot of this account is that it correctly diagnoses official accounts in line with EXTENSIONAL ADEQUACY.

A bonus feature of this account is its *relativisation*. Consider, for instance, the fact that the National Security Agency has conducted widespread surveillance on the American populace. Before 2013—the year when this was exposed to the general public and endorsed by the relevant epistemic authorities (journalists, lawyers, judges, etc.)—this would have been considered a conspiracy theory, but once it becomes a matter of historical record, we no longer categorise it such. This is correctly diagnosed on Harris' account as after 2013 it no longer conflicts with the relevant authorities' claims whereas pre-exposure, it would have conflicted (as there was not available evidence in support of it). So, the CONTRA EPISTEMIC AUTHORITIES DEFINITION has temporal relativisation in its characterising of conspiracy theory or not. Note further that it also has *geographical* relativisation insofar as the relevant epistemic authorities in one location might be different from the authorities in another. Thus the same explanation of events can conflict with the claims of a relevant epistemic authority in one location but not another, meaning the same explanation of an event can be a conspiracy theory to some group of people but not others. This is a feature and not a bug as people in some country may propose a conspiracy theory that conflicts with the relevant epistemic authorities that the expert authorities in some other country actually endorse; it seems quite natural to consider the first group conspiracy theorists and the second group not.

So, given that the CONTRA EPISTEMIC AUTHORITIES DEFINITION meets both our desiderata *and* appears to do well on the supplementary EXTENSIONAL ADEQUACY front, should it be the account we endorse? I suggest no, for the reasons of the following two problems I will now outline.

5.3.2.4 Two problems

The first issue relates to the relativisation. In some cases, who the relevant epistemic authorities are is clear. In other cases, however, it is not, and, more worryingly, it is not clear that there is a satisfactory or easy principle with which to adjudicate this—and Harris does not give one. For instance, what if you are in one country, but follow another country's geopolitics and society more than your own. Suppose further that you are aware of both country's epistemic authorities *and* they are in conflict with each other. Which is the relevant epistemic authority for you? It is not clear. This is not some far-fetched case either—different areas will have different authorities that hold different views but the individuals are not epistemically isolated such that they are unaware of this. Moreover, making it purely geographical seems arbitrary; it would be odd to think who your epistemic authorities are changes when you visit another country for a holiday. At the very least, some explanation or principle to follow here would be helpful, and we have not, so far, been given one.

Nevertheless, we might think this problem is not particularly damning. Yes, the account might not be wholly informative in some cases but it is rare that any account in philosophy is perfect. I agree with this. However, the account I will put forward in the following section does not have this relativisation problem. Nor does it have the following consequence, and this is one that I take to be far more problematic. In short, CONTRA EPISTEMIC AUTHORITIES DEFINITION implies that epistemic authorities cannot propagate conspiracy theories. This is especially troubling for two reasons. First, it is empirically and extensionally mistaken insofar as, for instance, in autocracies, the relevant epistemic authorities will often be the ones conspiring. Take, for example, leaders of the nuclear power plant at Chernobyl conspiring to cover up the disaster that occurred by blaming it on the mistakes of a few engineers at the plant who then tried to hide it. This is intuitively a conspiracy theory from the party yet they are presumably the epistemic authorities (i.e., credentialed individuals) so they cannot conflict with their own claims. Secondly, getting these sorts of cases wrong is particularly damning because we might think these are the conspiracy theories we care most about insofar as the epistemic authorities' own conspiracy theory propagation is likely to be far more dangerous than your average conspiracy theory engendered by the laity. Thus, these cases make for a substantial mark against the CONTRA EPISTEMIC AUTHORITIES DEFINITION's card.

Prima facie, however, it may appear that Harris has a nice response ready and waiting thanks to his account of epistemic authority allowing for well- and poorly functioning systems. The thought goes: any of these sorts of counterexamples where the relevant epistemic authorities are the ones putting forward conspiracy theories must be the product of dysfunctional systems (granted the plausible assumption that the experts in well-functioning systems tend not to conspire in this way), and so, we could claim, are not bona fide experts. Therefore, they lose their categorisation of

"relevant epistemic authorities" and we index that expertise elsewhere meaning that it is possible for their claims to conflict with the actual authorities.

Despite at first glance looking plausible, this response is nevertheless ultimately unsuccessful. Notice that, in going down this route, we remove the "relevant epistemic authority" categorisation whenever they go wrong. Therefore, the authorities referenced in the CONTRA EPISTEMIC AUTHORITIES DEFINITION will always be epistemically good. This has the upshot of making the definition a pejorative one because as soon as apparent authorities go wrong, then they were just never proper experts in the first place. This account then just collapses into an epistemically demarcated epistemic authority (a la Zagzebski (2012) and Goldman (1999)) which Harris specifically wants to avoid for the reasons outlined above.[35] Therefore, these counterexamples remain.

So, the CONTRA EPISTEMIC AUTHORITIES DEFINITION has two distinct issues facing it. Despite this, I do think the account is much preferable to the MINIMALIST DEFINITION and, perhaps, in the grand scheme of things, these issues are not ruinous. Nevertheless, as I will now demonstrate, my novel account maintains all of the upside of Harris', with none of the downsides.

5.4 A NOVEL ACCOUNT OF CONSPIRACY THEORIES

In this section, I outline and defend my novel account of conspiracy theories, the POSITION TO KNOW DEFINITION:

> POSITION TO KNOW DEFINITION
> A theory, t, is a conspiracy theory for an agent, A, if and only if:
>
> 1. t is an explanation of events that alleges a conspiracy, and
> 2. A is not in a position to know that t.

Following the CONTRA EPISTEMIC AUTHORITIES DEFINITION, I am taking the MINIMALIST DEFINITION as a necessary condition on my account, with the added caveat that I am understanding conspiracy here to be more than one agent trying to minimise awareness of some plan to bring about a somewhat *sinister* end. The addition of "sinister" serves to eliminate any "conspiracies of goodness" as Dentith calls them (2014, 49–50) while maintaining the brunt of the MINIMALIST DEFINITION—viz., all the official stories would still come out as conspiracy theories without further conditions. The real meat of my account is in the second condition, where it employs the technical epistemological term of the *position to know*. It to this we now turn.

35 Going this route would also presumably trivially entail the irrationality of conspiracy theories, failing our desiderata.

5.4.1 The Position to Know

In Christopher Willard-Kyle's pithy slogan, "The position to know is something like knowledge minus belief." (Willard-Kyle 2020, 329)[36] There are two key features of the position to know: how it *modalises* and how it *depsychologises*. Let us start with the first.

"One is in a position to know p if one could know it under the right circumstances." (330) Crucially, one does not need to know that p in order to be in a position to know that p. Instead, what arbitrates whether one is in a position to know that p is whether there is a possible world where someone *does know* that p while maintaining the same epistemic position—i.e., the same evidence, the same truth or falsity of the target proposition, and so on. So, to give an example, in the actual world, A is not in a position to know what year the Forth Road Bridge was constructed despite the fact that he could easily go on Wikipedia and check it right now. Moreover, the fact that there is some nearby possible where A *has* checked the Wikipedia and so knows it was constructed in 1964 still does not mean the A in the actual world is in a position to know. Why? Because the A in the nearby world is not in the same epistemic position as the A in the actual world—they have the extra Wikipedia evidence. The general idea is that, when in a position to know, the agent already has everything they need in order *to know*, they just need to take that "belief" step, harkening back to Willard-Kyle's slogan.

Turning now to the depsychologisation; what is paramount is the epistemic position itself, "not what use the agent has or is in fact psychologically capable of seeing from it." (330) Arbitrary facts about the agent's mental states do not impinge upon whether they are in a position to know that p—this will prove especially important considering our focus here is on conspiracy theories, theorists, and theorising. Consider the following example; suppose some agent B has a keen interest in space exploration and has read nigh on all there is to know about the various Apollo and Luna missions. Despite this, due to some facts about their mental states, B is simply psychologically incapable of believing that the moon landing is real (perhaps they are suffering from acute selenophobia despite their astronomical interests). In other words, B is unable to know that the moon landing is real. Nevertheless, B *is* in a position to know that the moon landing really happened because what dictates this is the *epistemic position*, not the psychological realities of the agent. B has read all about Apollo 11 and so their epistemic position regarding the moon landing is knowledge-conducive regardless of their selenophobia which is preventing them from coming to

36 The following exegesis is drawn primarily from Willard-Kyle (2020). This not to say that it is *the* analysis of the position to know but it is more than adequate for my purposes here. Indeed, Willard-Kyle has since updated his position in Kearl and Willard-Kyle (forthcoming) although I take everything said here to be commensurate with that recent work. Also see Williamson (2000) and Rosenkranz (2007) for some classic works on the position to know, and Simion (2024b) for the key contemporary research.

know. If B could just take that belief step, they would know. As Willard-Kyle puts it, "*Someone* in the agent's epistemic situation could know even if the agent himself could not: someone with a different psychological profile." (330) When determining whether someone is in a position to know some p, the epistemic position is held fixed, while the psychological states can be shifted. In light of the discussion so far, here is Willard-Kyle's gloss:

> POSITION TO KNOW:
> S is in a position to know that p iff S could know that p given their actual epistemic position. (330)

The exegesis is mostly complete. It is important, however, to first put some key constraints on the epistemic position in play. Here are three principles I will adopt from Willard-Kyle (2020, 331) that will be useful for my purposes here:

> *Position to Know to Truth*:
> If S is in a position to know that p then p is true.
>
> *Knows to Position to Know*:
> If S knows that p then S is in a position to know that p.
>
> *Anti-Collapse*:
> For some S and for some p, it is possible that S is in a position to know that p but does not know that p.

The first is a simple factivity constraint—one cannot be in a position to know something false. This is intuitive; I will never be in a position to know that Glasgow is the capital city of Scotland because Glasgow is not the capital city of Scotland. The second clarifies that knowing that p entails being in a position to know that p. This is roughly analogous to the principle of *actuality* entailing *possibility*. Finally, the third says, "in effect, there is no general inference from an agent's not knowing p to the agent's not being in a position to know that p." (331) The position to know does not collapse into knowing (it would be quite pointless if it did).

This completes the discussion of the position to know. It should now be clear what it is for an agent to not be in a position to know that t. So, I am now in a position to *show* that the POSITION TO KNOW DEFINITION meets both our desiderata while also meeting EXTENSIONAL ADEQUACY.

5.4.2 Evaluating the account

5.4.2.1 Extensional Adequacy

My novel definition does extremely well on the EXTENSIONAL ADEQUACY front. This is because it is suitably externalist in some respects while internalist in others. The factivity of the position to know prevents pure subjectivity on behalf of the relevant

agent, while the focus on the agent's own perspective prevents a purely objective standpoint on (psychologically relevant) conspiracy theories.

Let us start by considering some simple paradigm cases of conspiracy theories such as Covid-19 was spread by 5G towers, climate change is a deep-state hoax, or the world is run by interdimensional lizard people. These are easily dealt with by the *Position to Know to Truth* principle: the position to know is factive, the given conspiracy theories are false, therefore all agents are trivially not in a position to know them (it is, in fact, impossible for them to be in a position to know them). Thus, they are explanations of events that allege conspiracies and any given agent is not in a position to know them, so they are correctly diagnosed as conspiracy theories.

We can now turn to the cases that have featured in our discussion of the MINIMALIST DEFINITION and the CONTRA EPISTEMIC AUTHORITIES DEFINITION; we had official stories counting as conspiracy theories, and epistemic authorities themselves propagating conspiracy theories. Starting with the official accounts, such as Al Qaeda perpetrating 9/11 or that the Nixon administration tried to cover up their involvement in the break-in to the Watergate complex in 1972, recall that the MINIMALIST DEFINITION diagnosed these *as* conspiracy theories, while the CONTRA EPISTEMIC AUTHORITIES DEFINITION (correctly, at least per EXTENSIONAL ADEQUACY) did not. My account similarly gives an accurate diagnosis. For the average person, who (pace Pigden) I take *not* to be what we would usually call a "conspiracy theorist", they will likely already *know* that Al Qaeda perpetrated 9/11 and that Nixon was involved in Watergate. Therefore, per the *Knows to Position to Know* principle, such agents are also in a position to know these explanations that allege conspiracies, therefore they do not meet the second condition of the POSITION TO KNOW DEFINITION and so these official stories are correctly not counted as conspiracy theories for them.[37]

What about agents who, say, believe the "9/11 was an inside job" conspiracy theory? We might think that my account says that, for them, the official account is a conspiracy theory, which would be a clear mistake. Well, it depends what their epistemic position is. Conspiracy theorists are often described as "falling down the rabbit hole", with the idea being that they once believed all the true claims (that vaccines are safe and effective, that Al Qaeda perpetrated 9/11, etc.) but then starting getting bad information (usually from social media) such that they began to believe all sorts of conspiracy theories. 9/11 being an inside job will always come out as a conspiracy theory on my account because of the factivity constraint. Importantly, the official story *will not* for anyone fallen down the rabbit hole as just mentioned. This is because they are still in a position to know that 9/11 was perpetrated by Al Qaeda, they just currently do not (or even *cannot* if their pathology is such) know it because of various new beliefs they have caused by bad social media information. Recall that

37 Even if the agents are unduly reticent, anxious, or insouciant, they are still in a position to know thanks to the depsychologising.

the position to know *depsychologises*; the fact that they currently have a pathology that is preventing them from believing the established fact of the matter crucially does not mean they are not in a position to know it.

In fact, the only way I think we can arrive at official stories being conspiracy theories for some agent is if they are raised in such epistemic isolation that they genuinely have no evidence for Al Qaeda's responsibility for 9/11, and, in their (say) cult, they are told that this is a conspiracy theory put forward by the US government. In this case, yes, they are not in a position to know the official story of 9/11 and so it would come out as a conspiracy theory, but I actually think this is an *accurate* diagnosis. For such a benighted cogniser, it does not seem problematic at all that this is countenanced as a conspiracy theory for them—it strikes me as rather plausible that theories I treat as orthodox could count as conspiracy theories to someone in a cult or some isolated, autocratic state. What is important is that in more standard cases, where agents *are not* so epistemically isolated, the official account will not be characterised as a conspiracy theory.

For the epistemic authorities propagating conspiracy theories, we get accurate diagnoses as well. The source of the conspiracy theory is irrelevant in the POSITION TO KNOW DEFINITION unlike the CONTRA EPISTEMIC AUTHORITIES DEFINITION. When the epistemic authorities at Chernobyl attempted to cover up the disaster by alleging conspiracy, these are correctly counted as conspiracy theories. Again this is handled by the factivity constraint: the authorities were obviously not in a position to know the story they were spreading because it was false. This connects to a related benefit that those partaking in a given conspiracy will never count as conspiracy theorists on my account; they *know* the conspiracy is true, thus are in a position to know it, and so will never meet the second condition.

I noted a distinct benefit (and problem) of the CONTRA EPISTEMIC AUTHORITIES DEFINITION was its temporal and geographical relativisation of conspiracy theory designation. My account gets the same benefits without the problem. For instance, were someone pre-1973 to have believed in the Nixon administration's involvement in the Watergate scandal, then they would be believing an allegation of conspiracy without being in the position to know it (they would lack evidence as the evidence had not come out yet), thus they would be countenanced as believing in a conspiracy theory—an accurate diagnosis. Once the scandal was exposed, they would know it, and so it would no longer be a conspiracy theory for them as it would fail the second condition. Notably, for the journalists who exposed it, they would have known it while the rest of the world was not in a position to, generating the correct result that for anyone *but* them, it was a conspiracy theory.

The geographical relativisation comes from whomever the agent in play is, and their epistemic position will often be dictated by the circumstances of their location—i.e., the cult case from above. Crucially, however, my POSITION TO KNOW DEFINI-

TION does not have the problem of arbitrary relativisation (recall, I worried that it could be difficult to see who makes up some agents' epistemic authorities). The relativisation is always to a specific agent and their epistemic position so it makes no difference what their exact location is, or whether they have gone on holiday to a place with different authorities. So, the problems that plagued the CONTRA EPISTEMIC AUTHORITIES DEFINITION are not present, while all the benefits of relativisation are maintained.

In sum then, my novel POSITION TO KNOW DEFINITION meets EXTENSIONAL ADEQUACY in a way that none of the other accounts can. Nevertheless, this was a supplementary consideration to be used as a tie-breaker. We need to now see if my account can meet the two key desiderata.

5.4.2.2 No Trivial (Ir)rationality

My account straightforwardly meets NO TRIVIAL (IR)RATIONALITY. Similar to the discussion of Harris' account, it obviously does not trivially entail the *rationality* of conspiracy theories and more leans towards a negative conception. Nevertheless, it also does not trivially entail *irrationality* either. Just because one is not in a position to know something does not mean it is *unjustified*. One will not have been in a position to know for some justified false belief (say their reliable, trustworthy tells them p but is wrong on this occasion), but that does not mean it was *irrational* for them to believe. So, not being in the position to know does not guarantee anything about rationality—at most it merely makes it more *unlikely* that the belief will be rational.

5.4.2.3 No View Entailment

Accordingly then, the POSITION TO KNOW DEFINITION meets the second desideratum too as it clearly entails neither GENERALISM$_b$ nor PARTICULARISM$_b$. It does not say belief in conspiracy theories is inherently problematic because, as just noted, one can be justified in a belief that one was never in a position to know.[38] It also clearly does not trivially entail that there is *nothing* inherently problematic about conspiracy theory belief; not being in a position to know is still somewhat of an epistemic flaw. In this respect, then, my account is clearly one of GENERALISM$_t$ insofar as it says all conspiracy theories can be assessed as a class by this sans position to know. Importantly, however, it is also entirely commensurate with a PARTICULARISM$_t$ viewpoint as well because its relativisation to a specific agent ensures each case is evaluated individually. Thus, my account uniquely treads a careful path between GENERALISM and PARTICULARISM in a way that no other account does, giving proper scope for a genuinely mid-way starting point in the philosophy of conspiracy theories.

38 Of course, some would disagree with this—i.e., K=J proponents such as Williamson (n.d.). Nevertheless, one can employ the position to know apparatus without going down this route.

5.4.2.4 The Relativisation Objection

While I have highlighted my novel account's relativisation to an single agent as an asset of the view, one might object that this is a rather drastic move that requires special justification. Prima facie, I think this concern is an understandable one, but upon closer look I believe that this relativisation is not such a dramatic shift after all. There are two main reasons for this, one conceptual, and one based on the extant literature.

Starting with the former, the term "conspiracy theory" itself always carries along with it some relativisation—or at least it *ought to*. This is because, in natural language, it would be very strange to consider someone who is actually part of a conspiracy theory (a *conspirator*) as a conspiracy theorist or someone who believes in a conspiracy theory when they are precisely one of the individuals involved. Therefore, any talk of conspiracy theory will always include some relativisation from step one—unless one wants to bite the rather odd bullet of considering someone a conspiracy theorist for knowing that they took part in a conspiracy.[39]

Secondly, almost every account in the extant literature involves some kind of relativisation. For instance, Harris' CONTRA EPISTEMIC AUTHORITIES DEFINITION, or the many *counter-official-story* accounts in the literature (e.g., Hagen (2022), Brooks (2023), and Coady (2012)). In fact, I imagine this is because most scholars (tacitly or otherwise) acknowledge the point I just made above; "conspiracy theory" is a term that very naturally relativises, hence most reasonable accounts involve some way of dealing with this. Indeed, the only account in the literature that is absolute in its characterisations is the MINIMALIST DEFINITION, and a huge portion of this chapter has been spent carefully laying out the extremely unintuitive judgements it makes across the board—this is often *because* it is incapable of any relativisation. And while my account may be the only one (so far) that relativises to an individual agent, the fact that many relativise to groups of agents suggests that my move is more pedestrian than it may initially seem—and it avoids the issues I earlier outlined with deciding whom we relativise to—thus I do not think this worry is a particularly concerning one.

5.5 CONCLUDING REMARKS

In this chapter, I began by outlining the problematic impasse that lies at the heart of the philosophy of conspiracy theories involving the battle between GENERALISM and PARTICULARISM and the definitions wielded therein. This chapter made for a discussion of a sizeable chunk of the entire literature and (what I take to be) the central issues. I drew out two theoretical desiderata, NO TRIVIAL (IR)RATIONALITY and NO VIEW ENTAILMENT, both of which had uncharitable versions employed by the particularist wing of the literature, thus ensuring that even they had to be beholden to

39 Note that the POSITION TO KNOW DEFINITION avoids this consequence.

them. Using these desiderata, I argued for two rather surprising conclusions: first, that the darling of the literature, the MINIMALIST DEFINITION fails on both grounds that were supposed to vindicate it, while a generalist proponent from Harris, the CONTRA EPISTEMIC AUTHORITIES DEFINITION, actually did rather well. Nevertheless, I highlighted a few outlying issues with the account and thus put forward my novel POSITION TO KNOW DEFINITION.

I showed how it is more extensionally adequate than any other accounts on the market—even the surprisingly adequate CONTRA EPISTEMIC AUTHORITIES DEFINITION—while clearly meeting the two desiderata outlined from the extant literature. Moreover, I noted that the POSITION TO KNOW DEFINITION nicely walks the fine line between GENERALISM and PARTICULARISM. I then answered a worry about its relativisation and argued that such a fear is unfounded. As a closing note on this novel account, I want to add that I think it captures the "unofficialness" of conspiracy theories in a satisfying way. This is an aspect that many accounts (again, the aforementioned contra authorities/official stories accounts) tried to capture, but primarily through social or political means. My account captures the unofficialness through *epistemic* means, which makes it more robust and consistent in its characterisations. I take this to be a nice extra virtue of the account, among its many other, more obvious assets.

References

Aalberg, Toril and Jenssen, Anders Todal. 2007. "Do Television Debates in Multiparty Systems affect Viewers? A Quasi- experimental Study with First-time Voters." *Scandinavian Political Studies* 30 (1): 115–135. https://doi.org/10.1111/j.1467-9477.2007.00175.x.

Ahlstrom-Vij, Kristoffer. 2013. *Epistemic Paternalism.* Palgrave Macmillan UK. https://doi.org/10.1057/9781137313171.

Aikin, Scott. 2010. *Epistemology and the regress problem.* Routledge.

Aird, Rory. 2023. "A puzzle of epistemic paternalism." *Philosophical Psychology* 36 (5): 1011–1029. https://doi.org/10.1080/09515089.2022.2146490.

Ali, Shiza, Saeed, Mohammad Hammas, Aldreabi, Esraa, Blackburn, Jeremy, De Cristofaro, Emiliano, Zannettou, Savvas, and Stringhini, Gianluca. 2021. "Understanding the Effect of Deplatforming on Social Networks." In *13th ACM Web Science Conference 2021,* 187–195. WebSci '21. Association for Computing Machinery. https://doi.org/10.1145/3447535.3462637.

Allen, Jennifer, Watts, Duncan J., and Rand, David G. 2024. "Quantifying the impact of misinformation and vaccine-skeptical content on Facebook." *Science* 384 (6699): eadk3451. https://doi.org/10.1126/science.adk3451.

Arbuckle, J. G., Hobbs, J., Loy, A., Morton, L. W., Prokopy, L. S., and Tyndall, J. 2014. "Understanding Corn Belt farmer perspectives on climate change to inform engagement strategies for adaptation and mitigation." *Journal of Soil and Water Conservation* 69 (6): 505–516. https://doi.org/10.2489/jswc.69.6.505.

Baksi, Catherine. 2019. "Lady Chatterley's legal case: how the book changed the meaning of obscene." *The Guardian,* https://www.theguardian.com/law/2019/aug/01/lady-chatterleys-legal-case-how-the-book-changed-the-meaning-of-obscene.

Ballew, Matthew T., Leiserowitz, Anthony, Roser-Renouf, Connie, Rosenthal, Seth A., Kotcher, John E., Marlon, Jennifer R., Lyon, Erik, Goldberg, Matthew H., and Maibach, Edward W. 2019. "Climate change in the American mind: Data, tools, and trends." *Environment: Science and Policy for Sustainable Development* 61 (3): 4–18. https://doi.org/10.1080/00139157.2019.1589300.

Basham, Lee. 2018a. "Conspiracy theory particularism, both moral and epistemic, versus generalism." In *Taking conspiracy theories seriously,* edited by M R. X. Dentith, 39–58. Collective studies in knowledge and society. Rowman & Littlefield International.

Basham, Lee. 2018b. "Joining the conspiracy." *Argumenta* 3 (2): 271–290.

Basham, Lee and Dentith, M R. X. 2016. "Social Science's Conspiracy-Theory Panic: Now They Want to Cure Everyone." *Social Epistemology Review and Reply Collective*, 12–19. http://wp.me/p1Bfg0-3fi.

Battaly, Heather. 2018a. "Can Closed-mindedness be an Intellectual Virtue?" *Royal Institute of Philosophy Supplement* 84:23–45. https://doi.org/10.1017/S1358246118000053X.

———. 2018b. "CLOSED-MINDEDNESS AND DOGMATISM." *Episteme* 15 (3): 261–282. https://doi.org/10.1017/epi.2018.22.

———. 2021. "Engaging closed-mindedly with your polluted media feed." In *The Routledge Handbook of Political Epistemology*, 1st ed., edited by Michael Hannon and Jeroen de Ridder, 312–324. Routledge. https://doi.org/10.4324/9780429326769-38.

Beaver, David and Stanley, Jason. 2023. *The Politics of Language.* Princeton University Press. https://doi.org/10.1515/9780691242743.

Benton, Matthew A. 2011. "Two more for the knowledge account of assertion." *Analysis* 71 (4): 684–687. https://doi.org/10.1093/analys/anr085.

———. 2012. "Assertion, knowledge and predictions." *Analysis* 72 (1): 102–105.

———. 2018. "Lying, Belief, and Knowledge." In *The Oxford Handbook of Lying*, edited by Jörg Meibauer, 120–133. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198736578.013.9.

Benton, Matthew A. and van Elswyk, Peter. 2020. "Hedged Assertion." In *The Oxford Handbook of Assertion*, edited by Sanford C. Goldberg, 244–263. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190675233.013.11.

Bernecker, Sven, Flowerree, Amy K., and Grundmann, Thomas. 2021. *The epistemology of fake news.* Oxford University Press.

Betz-Richman, Noah. 2022. "Lying, hedging, and the norms of assertion." *Synthese* 200 (2): 176. https://doi.org/10.1007/s11229-022-03644-8.

Bond, Shannon. 2024. "This is what Russian propaganda looks like in 2024." *NPR*, https://www.npr.org/2024/06/06/g-s1-2965/russia-propaganda-deepfakes-sham-websites-social-media-ukraine.

Bondy, Patrick. 2020. "Epistemic Paternalism and Epistemic Normativity." In *Epistemic Paternalism: Conceptions, Justifications and Implications*, edited by Guy Axtell and Amiel Bernal, 141–154. Rowman & Littlefield Publishers.

Boudry, Maarten. 2022. "Why We Should Be Suspicious of Conspiracy Theories: A Novel Demarcation Problem." *Episteme*, 1–21. https://doi.org/10.1017/epi.2022.34.

Boudry, Maarten and Napolitano, M. Guilia. 2023. "Why We Should Stop Talking about Generalism and Particularism: Moving the Debate on Conspiracy Theories Forward." *Social Epistemology Review and Reply Collective* 12 (9): 22–26.

Brew, Simon. 2019. "The Myth of A Clockwork Orange's Ban." *Den of Geek*, https://www.denofgeek.com/movies/the-myth-of-a-clockwork-orange-s-ban/.

Broncano-Berrocal, Fernando. 2020. "Epistemic Care and Epistemic Paternalism." In *Epistemic Paternalism: Conceptions, Justifications and Implications*, edited by Amiel Bernal and Guy Axtell, 169–82. Rowman & Littlefield Publishers.

Brooks, Patrick. 2023. "On the origin of conspiracy theories." *Philosophical Studies* 180 (12): 3279–3299. ISSN: 1573-0883. https://doi.org/10.1007/s11098-023-02040-3.

Brown, Étienne. 2021. "Regulating the spread of online misinformation." In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 214–225. Routledge.

Brown, Penelope and Levinson, Stephen C. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Buenting, Joel and Taylor, Jason. 2010. "Conspiracy Theories and Fortuitous Data." *Philosophy of the Social Sciences* 40 (4): 567–578. https://doi.org/10.1177/0048393109350750.

Bullock, Emma C. 2018. "Knowing and Not-Knowing For Your Own Good: The Limits of Epistemic Paternalism." *Journal of Applied Philosophy* 35 (2): 433–447. https://doi.org/10.1111/japp.12220.

Carson, Thomas L. 2010. *Lying and Deception: Theory and Practice*. Oxford University Press.

Carter, J. Adam. 2020. "On behalf of a bi-level account of trust." *Philosophical Studies* 177 (8): 2299–2322. https://doi.org/10.1007/s11098-019-01311-2.

———. 2022. "Trust as performance." *Philosophical Issues* 32 (1): 120–147. https://doi.org/10.1111/phis.12214.

———. 2023a. *Stratified virtue epistemology: A defence*. Cambridge University Press.

———. 2023b. "Trust and trustworthiness." *Philosophy and Phenomenological Research* 107 (2): 377–394. https://doi.org/10.1111/phpr.12918.

———. 2024. *A telic theory of trust*. Oxford University Press.

Carter, J. Adam and Kallestrup, Jesper. Forthcoming. "Virtuous Deferral." *Noûs*, 1–37.

Carter, J. Adam and Willard-Kyle, Christopher. Forthcoming. "Virtue Epistemology for the Zetetic Turn." *MIND*, 1–18.

Cassam, Quassim. 2019a. *Conspiracy theories*. Think. Polity Press.

———. 2019b. *Vices of the mind: from the intellectual to the political*. First edition. Oxford University Press.

Castro, Clinton, Pham, Adam, and Rubel, Alan. 2020. "Epistemic Paternalism Online." In *Epistemic Paternalism: Conceptions, Justifications and Implications*, edited by Amiel Bernal and Guy Axtell, 29–44. Rowman & Littlefield Publishers.

Chantler-Hicks, Lydia. 2024. "How online lies and misinformation fuelled UK riots - and what really happened." *The Standard*, https://www.standard.co.uk/news/uk/southport-riots-violence-protests-far-right-tommy-robinson-edl-nigel-farage-reform-b1174829.html.

Chiou, Lesley and Tucker, Catherine E. 2018. "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement." *SSRN Electronic Journal*, 1–35. ISSN: 1556-5068. https://doi.org/10.2139/ssrn.3209929.

Cionea, Ioana A., Piercy, Cameron W., and Carpenter, Christopher J. 2017. "A profile of arguing behaviors on Facebook." *Computers in Human Behavior* 76:438–449. https://doi.org/10.1016/j.chb.2017.08.009.

Coady, David. 2012. *What to believe now: Applying epistemology to contemporary issues.* John Wiley & Sons.

Cohen, Daniel H. 2005. "Arguments That Backfire." In *The Uses of Argument*, edited by D. Hitchcock and D. Farr, 58–65. OSSA.

Cohen, Gerald A. 2002. "Deeper into bullshit." In *The Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. The MIT Press. https://doi.org/10.7551/mitpress/2143.003.0015.

Cova, Florian. 2024. "What's Wrong with Bullshit." *Ergo an Open Access Journal of Philosophy* 11 (0). https://doi.org/10.3998/ergo.6162.

Croce, Michel. 2018. "Epistemic Paternalism and the Service Conception of Epistemic Authority: EPISTEMIC PATERNALISM AND EPISTEMIC AUTHORITY." *Metaphilosophy* 49 (3): 305–327. https://doi.org/10.1111/meta.12294.

———. 2020. "Epistemic Paternalism, Personal Sovereignty, and One's Own Good." In *Epistemic Paternalism: Conceptions, Justifications and Implications*, edited by Guy Axtell and Amiel Bernal, 155–168. Rowman & Littlefield Publishers.

Dallison, Paul. 2019. "Bid to prosecute Boris Johnson over Brexit bus claim thrown out by court." *POLITICO*, https://www.politico.eu/article/boris-johnson-brexit-bus/.

Dentith, M R. X. 2014. *The Philosophy of Conspiracy Theories.* Palgrave Macmillan UK. https://doi.org/10.1057/9781137363169.

———. 2016. "When Inferring to a Conspiracy might be the Best Explanation." *Social Epistemology* 30 (5-6): 572–591. https://doi.org/10.1080/02691728.2016.1172362.

———. 2018. "Expertise and Conspiracy Theories." *Social Epistemology* 32 (3): 196–208. https://doi.org/10.1080/02691728.2018.1440021.

———. 2019. "Conspiracy theories on the basis of the evidence." *Synthese* 196:2243–2261.

———. 2022. "Suspicious conspiracy theories." *Synthese* 200 (3): 243. https://doi.org/10.1007/s11229-022-03602-4.

———. 2023a. "Some Conspiracy Theories." *Social Epistemology* 37 (4): 522–534. https://doi.org/10.1080/02691728.2023.2173539.

———. 2023b. "The Future of the Philosophy of Conspiracy Theory: An Introduction to the Special Issue on Conspiracy Theory Theory." *Social Epistemology* 37 (4): 405–412. https://doi.org/10.1080/02691728.2023.2173538.

Dentith, M R. X. and Tsapos, Melina. 2024. "Why We Should Talk about Generalism and Particularism: A Reply to Boudry and Napolitano." *Social Epistemology Review and Reply Collective* 13 (10): 47–60. https://social-epistemology.com/2024/10/25/why-we-should-talk-about-generalism-and-particularism-a-reply-to-boudry-and-napolitano-m-r-x-dentith-and-melina-tsapos/.

DeRose, Keith. 2002. "Assertion, knowledge, and context." *The Philosophical Review* 111 (2): 167–203.

Doherty, Carroll and Kiley, Jocelyn. 2023. "A Look Back at How Fear and False Beliefs Bolstered U.S. Public Support for War in Iraq." *Pew Research Center*, https://www.pewresearch.org/politics/2023/03/14/a-look-back-at-how-fear-and-false-beliefs-bolstered-u-s-public-support-for-war-in-iraq/.

Duetz, J. C. M. 2023. "What Does It Mean for a Conspiracy Theory to Be a 'Theory'?" *Social Epistemology* 37 (4): 438–453. https://doi.org/10.1080/02691728.2023.2172697.

———. 2024. "Conspiracy Theories are Not Beliefs." *Erkenntnis* 89 (5): 2105–2119. https://doi.org/10.1007/s10670-022-00620-z.

Dutilh Novaes, Catarina. 2023. "VII—Can Arguments Change Minds?" *Proceedings of the Aristotelian Society* 123 (2): 173–198. https://doi.org/10.1093/arisoc/aoad006.

Elford, Gideon. 2023. "No Platforming and Academic Freedom." *Ergo: An Open Access Journal of Philosophy* 10 (29): 808–839. https://doi.org/10.3998/ergo.4659.

Ettensperger, Felix, Waldvogel, Thomas, Wagschal, Uwe, and Weishaupt, Samuel. 2023. "How to convince in a televised debate: the application of machine learning to analyze why viewers changed their winner perception during the 2021 German chancellor discussion." *Humanities and Social Sciences Communications* 10 (546): 1–16. https://doi.org/10.1057/s41599-023-02047-5.

European Commission. 2018. "Fake news and disinformation online - March 2018 - - Eurobarometer survey." *Eurobarometer*, https://europa.eu/eurobarometer/surveys/detail/2183.

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

Fallis, Don and Mathiesen, Kay. 2019. "Fake news is counterfeit news." *Inquiry*, 1–20. https://doi.org/10.1080/0020174X.2019.1688179.

Fantl, Jeremy. 2018. *The limitations of the open mind.* Oxford University Press.

Feldman, Richard. 2000. "The Ethics of Belief." *Philosophy and Phenomenological Research* 60 (3): 667–695. https://doi.org/10.2307/2653823.

———. 2008. "Modest Deontologism in Epistemology." *Synthese* 161 (3): 339–355. https://doi.org/10.1007/s11229-006-9088-y.

Frankfurt, Harry G. 1986. "On Bullshit." *Raritan Quarterly* 6 (2): 81–100.

Frankfurt, Harry G. 2002. "Reply to GA Cohen." *Contours of agency: Essays on themes from Harry Frankfurt*, 340–344.

———. 2005. *On Bullshit*. Princeton University Press. https://doi.org/10.1515/9781400826537.

Fraser, Bruce. 2010. "Pragmatic Competence: The Case of Hedging." In *New Approaches to Hedging*, edited by Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider, 15–34. BRILL.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Funk, Cary and Kennedy, Brian. 2020. "For Earth Day 2020, how Americans see climate change and the environment in 7 charts." *Pew Research Center*, https://pewrsr.ch/2UqQsOI.

Gadsby, Stephen. 2023. "Bad beliefs: automaticity, arationality, and intervention." *Philosophical Psychology* 36 (4): 778–791. https://doi.org/10.1080/09515089.2023.2173060.

Gelfert, Axel. 2018. "Fake News: A Definition." *Informal Logic* 38 (1): 84–117. https://doi.org/10.22329/il.v38i1.5068.

Gerken, Mikkel. 2020. "How to balance Balanced Reporting and Reliable Reporting." *Philosophical Studies* 177 (10): 3117–3142. https://doi.org/10.1007/s11098-019-01362-5.

———. 2022. *Scientific Testimony: Its roles in science and society*. Oxford University Press.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23 (6): 121–123. https://doi.org/10.1093/analys/23.6.121.

Gibbons, Adam F. 2023. "Bullshit in Politics Pays." *Episteme*, 1–21. https://doi.org/10.1017/epi.2023.3.

Gilbert, David. 2021. "Facebook Knew It Was Fueling QAnon." *Vice*, https://www.vice.com/en/article/facebooks-algorithm-spread-qanon-content-to-new-users/?utm_source=motherboard_twitter.

Gjelsvik, Olav. 2018. "Bullshit Production." In *Lying: Language, Knowledge, Ethics, and Politics*, edited by Eliot Michaelson and Andreas Stokke, 129–142. Oxford University Press. https://doi.org/10.1093/oso/9780198743965.003.0007.

Godden, David. 2020. "Epistemic autonomy, epistemic paternalism, and blindspots of reason." In *Epistemic paternalism: Conceptions, justifications, and implications*, edited by Amiel Bernal and Guy Axtell, 183–200. Rowman & Littlefield Publishers.

Goldberg, Jeffrey. 2020. "Why Obama Fears for Our Democracy?" *The Atlantic*, https://www.theatlantic.com/ideas/archive/2020/11/why-obama-fears-for-our-democracy/617087/.

Goldberg, Sanford C. 2017. "Should have known." *Synthese* 194 (8): 2863–2894. https://doi.org/10.1007/s11229-015-0662-z.

———. 2018. *To the best of our knowledge: Social expectations and epistemic normativity.* Oxford University Press.

———. 2020. *Conversational pressure: Normativity in speech exchanges.* Oxford University Press.

Goldman, Alvin I. 1988. "Strong and Weak Justification." *Philosophical Perspectives* 2:51–69. https://doi.org/10.2307/2214068.

———. 1991. "Epistemic Paternalism: Communication Control in Law and Society." *The Journal of Philosophy* 88 (3): 113. https://doi.org/10.2307/2026984.

———. 1999. *Knowledge in a social world.* Oxford University Press.

Gordon-Smith, Eleanor. 2019. *Stop being reasonable: how we really change our minds.* PublicAffairs.

Graeber, David. 2018. *Bullshit jobs: a theory.* Simon & Schuster.

Granieri, Ronald J. 2020. "The right needs to stop falsely claiming that the Nazis were socialists." *The Washington Post*, accessed March 31, 2025. https://www.washingtonpost.com/outlook/2020/02/05/right-needs-stop-falsely-claiming-that-nazis-were-socialists/.

Grice, Herbert Paul. 1978. "Further notes on logic and conversation." *Syntax and semantics* 9.

———. 1989. *Studies in the Way of Words.* Cambridge: Harvard University Press.

Griffiths, James. 2019. "World marks 30 years since Tiananmen massacre as China censors all mention." *CNN*, https://edition.cnn.com/2019/06/03/asia/tiananmen-june-4-china-censorship-intl.

Grundmann, Thomas. 2023. "Fake news: the case for a purely consumer-oriented explication." *Inquiry* 66 (10): 1758–1772. https://doi.org/10.1080/0020174X.2020.1813195.

Habgood-Coote, Joshua. 2023. "Deepfakes and the epistemic apocalypse." *Synthese* 201 (3): 103. https://doi.org/10.1007/s11229-023-04097-3.

Hagen, Kurtis. 2022. *Conspiracy theories and the failure of intellectual critique.* University of Michigan Press.

———. 2024. "Particularism as the Corrective to the Conventional Wisdom Regarding Conspiracy Theories." *Social Epistemology Review and Reply Collective*, 12–24. https://wp.me/p1Bfg0-9mo.

———. 2025. "Generalist Denialism and the Particularist Critique." *Social Epistemology Review and Reply Collective*, 35–45. https://wp.me/p1Bfg0-9yW.

Halpert, Madeline. 2025. "Measles outbreak in west Texas worsens due to vaccine scepticism." *BBC News*, https://www.bbc.com/news/articles/cwy7eyde3xeo.

Harris, Keith Raymond. 2018. "What's Epistemically Wrong with Conspiracy Theorising?" *Royal Institute of Philosophy Supplement* 84:235–257. https://doi.org/10.1017/S1358246118000619.

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

———. 2022. "Some problems with particularism." *Synthese* 200 (6): 447. https://doi.org/10.1007/s11229-022-03948-9.

———. 2023. "Conspiracy Theories, Populism, and Epistemic Autonomy." *Journal of the American Philosophical Association* 9 (1): 21–36. https://doi.org/10.1017/apa.2021.44.

———. 2024. *Misinformation, Content Moderation, and Epistemology: Protecting Knowledge.* Routledge.

Haslanger, Sally. 2012. *Resisting reality: Social construction and social critique.* Oxford University Press.

Hauswald, Rico. 2023. ""That's Just a Conspiracy Theory!": Relevant Alternatives, Dismissive Conversational Exercitives, and the Problem of Premature Conclusions." *Social Epistemology* 37 (4): 494–509. http://dx.doi.org/10.1080/02691728.2023.2172699.

Hazlett, Allan. 2020. "False Intellectual Humility." In *The Routledge Handbook of the Philosophy of Humility*, edited by Mark Alfano, Michael Patrick Lynch, and Alessandra Tanesini, 313–324. Routledge.

Hicks, Michael Townsen, Humphries, James, and Slater, Joe. 2024. "ChatGPT is bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hill, Scott. 2024. "Particularism and the Conventional Wisdom." *Social Epistemology Review and Reply Collective*, 44–51. https://wp.me/p1Bfg0-9og.

Innes, H. and Innes, M. 2021. "De-platforming disinformation: conspiracy theories and their control." *Information, Communication & Society* 0 (0): 1–19. https://doi.org/10.1080/1369118X.2021.1994631.

Jackson, Elizabeth. 2020. "Epistemic paternalism, epistemic permissivism, and standpoint epistemology." In *Epistemic Paternalism: Conceptions, Justifications and Implications*, edited by Amiel Bernal and Guy Axtell, 203–218. Rowman & Littlefield Publishers.

———. 2021. "What's Epistemic about Epistemic Paternalism?" In *Epistemic Autonomy*, edited by Kirk Lougheed and Jonathan Matheson, 132–150. Routledge.

Jhaver, Shagun, Boylston, Christian, Yang, Diyi, and Bruckman, Amy. 2021. "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 381:1–381:30. https://doi.org/10.1145/3479525.

John, Stephen. 2018. "Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty." *Social Epistemology* 32 (2): 75–87. https://doi.org/10.1080/02691728.2017.1410864.

Johnson, Andrew. 2010. "A New Take on Deceptive Advertising: Beyond Frankfurt's Analysis of 'BS'." *Business & Professional Ethics Journal* 29 (1/4): 5–32.

Johnson, Casey Rebecca. 2018. "Just Say 'No': Obligations to Voice Disagreement." *Royal Institute of Philosophy Supplements* 84:117–138. https://doi.org/10.1017/S1358246118000577.

Jong-Fast, Molly. 2022. "Owning the Libs Is the Only GOP Platform." *The Atlantic*, https://www.theatlantic.com/newsletters/archive/2022/01/owning-the-libs-is-the-only-gop-platform/676692/?utm_source=copy-link&utm_medium=social&utm_campaign=share.

Joshi, Hrishikesh. 2024. "The censor's burden." *Noûs*, https://doi.org/10.1111/nous.12534.

Kearl, Timothy and Willard-Kyle, Christopher. Forthcoming. "Epistemic cans." *Philosophy and Phenomenological Research*.

Keeley, Brian. 1999. "Of Conspiracy Theories." *Journal of Philosophy* 96 (3): 109–126. https://doi.org/10.2139/ssrn.1084585.

Kitsik, Eve. 2023. "Epistemic paternalism via conceptual engineering." *Journal of the American Philosophical Association* 9 (4): 616–635. https://doi.org/doi:10.1017/apa.2022.22.

Klieber, Anna. 2024. "Conversational silence, reconsidered." *Theoria* 90 (6): 652–668. https://doi.org/10.1111/theo.12566.

Lackey, Jennifer. 2018. "Silence and Objecting." In *Voicing Dissent: The Ethics and Epistemology of Making Disagreement Public*, 1st ed., edited by Casey Rebecca Johnson, 82–96. Routledge. https://doi.org/10.4324/9781315181189.

———. 2020a. "Epistemic Duties Regarding Others." In *Epistemic Duties*, edited by Kevin McCain and Scott Stapleford, 281–295. Routledge.

———. 2020b. "The Duty to Object." *Philosophy and Phenomenological Research* 101 (1): 35–60. https://doi.org/10.1111/phpr.12563.

———. 2021. "When Should We Disagree About Politics?" In *Political epistemology*, edited by Elizabeth Edenberg and Michael Hannon, 280–296. Oxford University Press.

Lakoff, George. 1973. "Hedges: A study in meaning criteria and the logic of fuzzy concepts." *Journal of Philosophical Logic* 2 (4). https://doi.org/10.1007/BF00262952.

Leiserowitz, Anthony, Maibach, Edward W., Roser-Renouf, Connie, Feinberg, Geoff, and Howe, Peter. 2013. "Climate change in the American mind: Americans' global warming beliefs and attitudes in April 2013." *Available at SSRN 2298705*, https://doi.org/10.2139/ssrn.2298705.

Levy, Neil. 2007. "Radically Socialized Knowledge and Conspiracy Theories." *Episteme* 4 (2): 181–192. https://doi.org/10.3366/epi.2007.4.2.181.

———. 2017. "The bad news about fake news." *Social epistemology review and reply collective* 6 (8): 20–36. http://wp.me/p1Bfg0-3GV.

Levy, Neil. 2019. "No-Platforming and Higher-Order Evidence, or Anti-Anti-No-Platforming." *Journal of the American Philosophical Association* 5 (4): 487–502. https://doi.org/10.1017/apa.2019.29.

———. 2021. *Bad Beliefs: Why They Happen to Good People.* 1st ed. Oxford University Press. https://doi.org/10.1093/oso/9780192895325.001.0001.

Lewis, David. 1996. "Elusive knowledge." *Australasian Journal of Philosophy* 74 (4): 549–567. https://doi.org/10.1080/00048409612347521.

Lyons, Silas. 2021. "Setting the record straight: Ad repeated lies about safety of COVID-19 vaccine." *Record Searchlight*, https://eu.redding.com/story/opinion/2021/09/25/setting-record-straight-ad-lies-safety-covid-19-vaccine/5850662001/.

Mackey, Robert and Lee, Micah. 2022. "Left-Wing Voices Are Silenced on Twitter as Far-Right Trolls Advise Elon Musk." *The Intercept*, https://theintercept.com/2022/11/29/elon-musk-twitter-andy-ngo-antifascist/.

Macleod, Christopher. 2021. "Mill on the Liberty of Thought and Discussion." In *The Oxford Handbook of Freedom of Speech*, edited by Adrienne Stone and Frederick Schauer, 2–19. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198827580.013.1.

Mahon, James Edwin. 2016. "The Definition of Lying and Deception." In *The Stanford Encyclopedia of Philosophy*, Winter 2016, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Mandik, Pete. 2007. "Shit Happens." *Episteme* 4 (2): 205–218. https://doi.org/10.3366/epi.2007.4.2.205.

Manne, Kate. 2017. *Down girl: The logic of misogyny.* Oxford University Press.

Mari, Silvia, Gil de Zúñiga, Homero, Suerdem, Ahmet, Hanke, Katja, Brown, Gary, Vilar, Roosevelt, Boer, Diana, and Bilewicz, Michal. 2021. "Conspiracy Theories and Institutional Trust: Examining the Role of Uncertainty Avoidance and Active Social Media Use." *Political Psychology* 43 (2): 277–296. https://doi.org/10.1111/pops.12754.

McCormick, Miriam Schleifer. 2023. "Engaging with "Fringe" Beliefs: Why, When, and How." *Episteme*, 1–16. https://doi.org/10.1017/epi.2023.33.

McCready, E. 2014. *Reliability in Pragmatics.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198702832.001.0001.

McIntyre, Lee. 2018. *Post-Truth.* MIT Press.

———. 2020. "Science denial, polarisation, and arrogance." In *Polarisation, Arrogance, and Dogmatism*, edited by Alessandra Tanesini and Michael P. Lynch, 193–211. Routledge.

———. 2021. *How to talk to a science denier: conversations with flat earthers, climate deniers, and others who defy reason.* MIT Press.

McKenna, Robin. 2020. "Persuasion and epistemic paternalism." In *Epistemic Paternalism: Conceptions, Justifications, and Implications*, edited by Amiel Bernal and Guy Axtell, 91–106. Rowman & Littlefield Publishers.

———. 2023. *Non-ideal epistemology*. Oxford University Press.

McMahon, Liv, Kleinman, Zoe, and Subramanian, Courtney. 2025. "Facebook and Instagram get rid of fact checkers." *BBC News*, https://www.bbc.co.uk/news/articles/cly74mpy8klo.

McMullen, Jane. 2022. "The audacious PR plot that seeded doubt about climate change." *BBC News*, https://www.bbc.com/news/science-environment-62225696.

Medvecky, Fabien. 2020. "Epistemic paternalism, science, and communication." In *Epistemic Paternalism: Conceptions, justifications and implications.*

Mill, John Stuart. 1859. *On Liberty*. Yale University Press.

Miller, Steve. 2001. "Public understanding of science at the crossroads." *Public Understanding of Science* 10 (1): 115–120. https://doi.org/10.3109/a036859.

Mitchell, Amy, Gottfried, Jeffrey, Stocking, Galen, Walker, Mason, and Fedeli, Sophia. 2019. "Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed." *Pew Research Center*, https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

Mittendorf, Will. 2023. "Conspiracy Theories and Democratic Legitimacy." *Social Epistemology* 37 (4): 481–493. https://doi.org/10.1080/02691728.2023.2172700.

Moberger, Victor. 2020. "Bullshit, Pseudoscience and Pseudophilosophy." *Theoria* 86 (5): 595–611. https://doi.org/10.1111/theo.12271.

Müller, Jan-Werner. 2024. "The US right keeps accusing Democrats of 'communism'. What does that even mean?" *The Guardian*, accessed March 31, 2025. https://www.theguardian.com/global/commentisfree/article/2024/sep/05/communism-meaning-republicans.

Munn, Luke. 2020. "Angry by design: toxic communication and technical architectures." *Humanities and Social Sciences Communications* 7 (1): 1–11. https://doi.org/10.1057/s41599-020-00550-7.

Napolitano, M. Giulia. 2021. "Conspiracy Theories and Evidential Self-Insulation." In *The Epistemology of Fake News*, 1st ed., edited by Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann, 82–106. Oxford University Press. https://doi.org/10.1093/oso/9780198863977.003.0005.

Napolitano, M. Giulia and Reuter, Kevin. 2023. "What is a Conspiracy Theory?" *Erkenntnis* 88 (5): 2035–2062. https://doi.org/10.1007/s10670-021-00441-6.

Nguyen, C. Thi. 2020. "ECHO CHAMBERS AND EPISTEMIC BUBBLES." *Episteme* 17 (2): 141–161. https://doi.org/10.1017/epi.2018.32.

Nisbet, Matthew C. and Scheufele, Dietram A. 2009. "What's next for science communication? Promising directions and lingering distractions." *American Journal of Botany* 96 (10): 1767–1778. https://doi.org/10.3732/ajb.0900041.

Nwokora, Zim and Brown, Lara M. 2017. "Narratives of a Race: How the Media Judged a Presidential Debate." *American Politics Research* 45 (1): 33–62. https://doi.org/10.1177/1532673X15614891.

Nyhan, Brendan and Reifler, Jason. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32 (2): 303–330. https://doi.org/10.1007/s11109-010-9112-2.

———. 2015. "Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information." *Vaccine* 33 (3): 459–464. https://doi.org/10.1016/j.vaccine.2014.11.017.

Oreskes, Naomi and Conway, Erik M. 2011. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming.* Bloomsbury Publishing USA.

Paglieri, Fabio. 2013. "Choosing to argue: Towards a theory of argumentative decisions." *Journal of Pragmatics*, Biases and constraints in communication: Argumentation, persuasion and manipulation, 59:153–163. https://doi.org/10.1016/j.pragma.2013.07.010.

Pennycook, Gordon and Rand, David G. 2022. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation." *Nature Communications* 13 (1). https://doi.org/10.1038/s41467-022-30073-5.

Peters, Uwe and Nottelmann, Nikolaj. 2021. "Weighing the Costs: The Epistemic Dilemma of No-Platforming." *Synthese* 199 (3-4): 7231–7253. https://doi.org/10.1007/s11229-021-03111-w.

Petrocelli, John V. 2018. "Antecedents of bullshitting." *Journal of Experimental Social Psychology* 76:249–258. https://doi.org/10.1016/j.jesp.2018.03.004.

Petrocelli, John V., Seta, Catherine E., and Seta, John J. 2023. "Lies and bullshit: The negative effects of misinformation grow stronger over time." *Applied Cognitive Psychology* 37 (2): 409–418. https://doi.org/10.1002/acp.4043.

Petrocelli, John V., Silverman, Haley E., and Shang, Samantha X. 2023. "Social perception and influence of lies vs. bullshit: a test of the insidious bullshit hypothesis." *Current Psychology* 42 (12): 9609–9617. https://doi.org/10.1007/s12144-021-02243-z.

Petrocelli, John V., Watson, Haley F., and Hirt, Edward R. 2020. "Self-Regulatory Aspects of Bullshitting and Bullshit Detection." *Social Psychology* 51 (4): 239–253. https://doi.org/10.1027/1864-9335/a000412.

Pierre, Joseph. 2020. "Mistrust and Misinformation: A Two-Component, Socio-Epistemic Model of Belief in Conspiracy Theories." *PsyArXiv*, https://doi.org/10.31234/osf.io/xhw52.

Pigden, Charles. 1995. "Popper Revisited, or What Is Wrong With Conspiracy Theories?" *Philosophy of the Social Sciences* 25 (1): 3–34. https://doi.org/10.1177/004839319502500101.

———. 2006. "Complots of mischief." In *Conspiracy theories: The philosophical debate*, edited by David Coady, 139–166.

———. 2007. "Conspiracy Theories and the Conventional Wisdom." *Episteme* 4 (2): 219–232. https://doi.org/10.3366/epi.2007.4.2.219.

———. 2016. "Are Conspiracy Theorists Epistemically Vicious?" In *A Companion to Applied Philosophy*, edited by Kasper Lippert-Rasmussen, Kimberley Brownlee, and David Coady, 120–132. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118869109.ch9.

———. 2018. "Conspiracy theories, deplorables, and defectibility: A reply to Patrick Stokes." In *Taking conspiracy theories seriously*, edited by Matthew R. X. Dentith, 203–215. Collective studies in knowledge and society. Rowman & Littlefield International.

Popper, Karl. 1966. *The open society and its enemies: The high tide of prophecy: Hegel, Marx and the aftermath.* Vol. 2. Routledge.

Pritchard, Duncan. 2013. "Epistemic paternalism and epistemic value." *Philosophical Inquiries* 1 (2): 9–37. https://doi.org/10.4454/philinq.v1i2.53.

Pummerer, Lotte, Böhm, Robert, Lilleholt, Lau, Winter, Kevin, Zettler, Ingo, and Sassenberg, Kai. 2021. "Conspiracy Theories and Their Societal Effects During the COVID-19 Pandemic." *Social Psychological and Personality Science* 13 (1): 49–59. https://doi.org/10.1177/19485506211000217.

Räikkä, Juha. 2018. "Conspiracies and conspiracy theories: An introduction." *Argumenta* 6:1–12. https://doi.org/10.23811/51.arg2017.rai.

Rauchfleisch, Adrian and Kaiser, Jonas. 2021. "Deplatforming the Far-right: An Analysis of YouTube and BitChute." *SSRN Scholarly Paper*, https://doi.org/10.2139/ssrn.3867818.

Rini, Regina. 2017. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27 (2): E–64. https://doi.org/10.1353/ken.2017.0025.

———. 2020. "Deepfakes and the Epistemic Backstop." *Philosopher's Imprint* 20 (24): 1–16. http://hdl.handle.net/2027/spo.3521354.0020.024.

Rosenblum, Nancy L. 2020. "Introduction: Paths to Witnessing, Ethics of Speaking Out." *Daedalus* 149 (4): 6–24. https://doi.org/10.1162/daed_e_01813.

Rosenkranz, Sven. 2007. "Agnosticism as a third stance." *Mind* 116 (461): 55–104. https://doi.org/10.1093/mind/fzm055.

Sah, Sunita, Moore, Don A., and MacCoun, Robert J. 2013. "Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness." *Organizational Behavior and Human Decision Processes* 121 (2): 246–255.

Salager-Meyer, Françoise. 1997. "I think that perhaps you should: A study of hedges in written scientific discourse." *Functional approaches to written text: Classroom applications* 1:127–143.

Satariano, Adam and Alba, Davey. 2020. "Burning Cell Towers, Out of Baseless Fear They Spread the Virus." *The New York Times*, https://www.nytimes.com/2020/04/10/technology/coronavirus-5g-uk.html.

Saul, Jennifer. 2021. "Someone Is Wrong on the Internet: Is There an Obligation to Correct False and Oppressive Speech on Social Media?" In *The Epistemology of Deceit in a Postdigital Era: Dupery by Design*, edited by Alison MacKenzie, Jennifer Rose, and Ibrar Bhatt, 139–157. Postdigital Science and Education. Springer International Publishing. https://doi.org/10.1007/978-3-030-72154-1_8.

Schaab, Janis David. 2022. "Conspiracy Theories and Rational Critique: A Kantian Procedural Approach." *Inquiry*, 1–30. https://doi.org/10.1080/0020174X.2022.2074883.

Schmid, Philipp and Betsch, Cornelia. 2019. "Effective strategies for rebutting science denialism in public discussions." *Nature Human Behaviour* 3 (9): 931–939. https://doi.org/10.1038/s41562-019-0632-4.

Schwemmer, Carsten. 2021. "The Limited Influence of Right-Wing Movements on Social Media User Engagement." *Social Media + Society* 7 (3): 20563051211041650. https://doi.org/10.1177/20563051211041650.

Shankland, R. S. 1964. "Michelson-Morley Experiment." *American Journal of Physics* 32 (1): 16–35. https://doi.org/10.1119/1.1970063.

Shen, Qinlan and Rosé, Carolyn P. 2022. "A Tale of Two Subreddits: Measuring the Impacts of Quarantines on Political Engagement on Reddit." *Proceedings of the International AAAI Conference on Web and Social Media* 16:932–943. https://doi.org/10.1609/icwsm.v16i1.19347.

Shields, Matthew. 2022. "Rethinking conspiracy theories." *Synthese* 200 (4): 331. https://doi.org/10.1007/s11229-022-03811-x.

———. 2023. "Conceptual Engineering, Conceptual Domination, and the Case of Conspiracy Theories." *Social Epistemology* 37 (4): 464–480. https://doi.org/10.1080/02691728.2023.2172696.

Simion, Mona. 2016. "Assertion: knowledge is enough." *Synthese* 193 (10): 3041–3056. https://doi.org/10.1007/s11229-015-0914-y.

———. 2021. "Blame as performance." *Synthese* 199 (3): 7595–7614. https://doi.org/10.1007/s11229-021-03130-7.

———. 2024a. "Knowledge and Disinformation." *Episteme* 21 (4): 1208–1219. https://doi.org/doi:10.1017/epi.2023.25.

———. 2024b. "Resistance to Evidence and the Duty to Believe." *Philosophy and Phenomenological Research* 108 (1): 203–216. https://doi.org/10.1111/phpr.12964.

Simion, Mona, Kelp, Christoph, and Ghijsen, Harmen. 2016. "Norms of Belief." *Philosophical Issues* 26 (1): 374–392. https://doi.org/10.1111/phis.12077.

Simis, Molly J., Madden, Haley, Cacciatore, Michael A., and Yeo, Sara K. 2016. "The lure of rationality: Why does the deficit model persist in science communication?" *Public Understanding of Science* 25 (4): 400–414. https://doi.org/10.1177/0963662516629749.

Simpson, Robert Mark. 2021. "Norms of Inquiry, Student-Led Learning, and Epistemic Paternalism." In *Epistemic Autonomy*, edited by Jonathan Matheson and Kirk Lougheed, 95–112. Routledge.

Simpson, Robert Mark and Srinivasan, Amia. 2018. "No platforming." In *Academic Freedom*, edited by Jennifer Lackey, 186–209. Oxford University Press.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–243. http://www.jstor.org/stable/2265052.

Sosa, Ernest. 2007. *A virtue epistemology: Apt belief and reflective knowledge, volume I.* Vol. 1. OUP Oxford.

———. 2010. "How Competence Matters in Epistemology." *Philosophical Perspectives* 24:465–475.

———. 2015. *Judgment and agency*. Oxford University Press, USA.

———. 2017. *Epistemology*. Princeton University Press. https://doi.org/10.1515/9781400883059.

———. 2021. *Epistemic explanations: A theory of telic normativity, and what it explains.* Oxford University Press.

Stalnaker, R. 1978. "Assertion." *Syntax and Semantics* 9.

Stamatiadis-Bréhier, Alexios. 2023. "Genealogical undermining for conspiracy theories." *Inquiry*, 1–27. https://doi.org/10.1080/0020174x.2023.2187449.

———. 2024. "The power of second-order conspiracies." *Inquiry*, 1–26. https://doi.org/10.1080/0020174x.2024.2375781.

Stanton, Zack. 2020. "You're Living in the Golden Age of Conspiracy Theories." *POLITICO*, https://www.politico.com/news/magazine/2020/06/17/conspiracy-theories-pandemic-trump-2020-election-coronavirus-326530.

Stokes, Patrick. 2018. "Conspiracy Theory and the Perils of Pure Particularism." In *Taking Conspiracy Theories Seriously*, edited by M R. X. Dentith, 25–37. Rowman & Littlefield International.

Stokke, Andreas. 2013a. "Lying and asserting." *The Journal of philosophy* 110 (1): 33–60.

———. 2013b. "Lying, Deceiving, and Misleading." *Philosophy Compass* 8 (4): 348–359. https://doi.org/https://doi.org/10.1111/phc3.12022.

Stokke, Andreas and Fallis, Don. 2017. "Bullshitting, Lying, and Indifference toward Truth." *Ergo, an Open Access Journal of Philosophy* 4. https://doi.org/10.3998/ergo.12405314.0004.010.

Suarez-Lledo, Victor and Alvarez-Galvez, Javier. 2021. "Prevalence of Health Misinformation on Social Media: Systematic Review." *Journal of Medical Internet Research* 23 (1): e17187. https://doi.org/10.2196/17187.

Sunstein, Cass R. 2002. "The Law of Group Polarization." *Journal of Political Philosophy* 10 (2): 175–195. https://doi.org/10.1111/1467-9760.00148.

Sylvan, Kurt L. 2020. "An Epistemic Nonconsequentialism." *The Philosophical Review* 129 (1): 1–51. https://doi.org/10.1215/00318108-7890455.

Tanesini, Alessandra. 2018. "Eloquent silences: Silence and dissent." In *Voicing Dissent: the ethics and epistemology of making disagreements public,* edited by Casey Rebecca Johnson, 109–128. Routledge.

Tappin, Ben M., Pennycook, Gordon, and Rand, David G. 2021. "Rethinking the link between cognitive sophistication and politically motivated reasoning." *Journal of Experimental Psychology. General* 150 (6): 1095–1114. https://doi.org/10.1037/xge0000974.

Tenney, Elizabeth R., MacCoun, Robert J., Spellman, Barbara A., and Hastie, Reid. 2007. "Calibration Trumps Confidence as a Basis for Witness Credibility." *Psychological Science* 18 (1): 46–50. https://doi.org/10.1111/j.1467-9280.2007.01847.x.

Tenney, Elizabeth R., Spellman, Barbara A., and MacCoun, Robert J. 2008. "The benefits of knowing what you know (and what you don't): How calibration affects credibility." *Journal of Experimental Social Psychology* 44 (5): 1368–1375.

Terzian, Giulia and Corbalán, M. Inés. 2021. "Our Epistemic Duties in Scenarios of Vaccine Mistrust." *International Journal of Philosophical Studies* 29 (4): 613–640. https://doi.org/10.1080/09672559.2021.1997399.

The Lancet Infectious Diseases. 2020. "The COVID-19 infodemic." *The Lancet* 20 (8). https://doi.org/10.1016/S1473-3099(20)30565-X.

Theel, Shauna, Greenberg, Max, and Robbins, Denise. 2013. "Media sowed doubt in coverage of UN climate report." *Media Matters for Americans. October* 10:2013. https://www.mediamatters.org/washington-post/study-media-sowed-doubt-coverage-un-climate-report.

Thorpe, Alistair, Fagerlin, Angela, Butler, Jorie, Stevens, Vanessa, Drews, Frank A., Shoemaker, Holly, Riddoch, Marian S., and Scherer, Laura D. 2022. "Communicating about COVID-19 vaccine development and safety." *PLOS ONE* 17 (8). https://doi.org/10.1371/journal.pone.0272426.

Titus, Lisa Miracchi and Carter, J. Adam. 2024. "What the tortoise should do: A knowledge-first virtue approach to the basing relation." *Noûs* 58 (2): 456–481. https://doi.org/10.1111/nous.12460.

Tranter, Bruce and Booth, Kate. 2015. "Scepticism in a changing climate: A cross-national study." *Global Environmental Change* 33:154–164. https://doi.org/10.1016/j.gloenvcha.2015.05.003.

Tsapos, Melina. 2023. "Who is a Conspiracy Theorist?" *Social Epistemology* 37 (4): 454–463. https://doi.org/10.1080/02691728.2023.2172695.

———. 2024a. "Betting on Conspiracy: A Decision Theoretic Account of the Rationality of Conspiracy Theory Belief." *Erkenntnis*, https://doi.org/10.1007/s10670-024-00785-9.

———. 2024b. "Should we worry about conspiracy theorists rejecting experts?" *Inquiry*, 1–21. https://doi.org/10.1080/0020174x.2024.2375774.

Tulin, Marina, Hameleers, Michael, de Vreese, Claes, Opgenhaffen, Michaël, and Wouters, Ferre. 2024. "Beyond Belief Correction: Effects of the Truth Sandwich on Perceptions of Fact-checkers and Verification Intentions." *Journalism Practice*, 1–20. https://doi.org/10.1080/17512786.2024.2311311.

United States Courts. *Federal Rules of Civil Procedure*. https://www.uscourts.gov/rules-policies/current-rules-practice-procedure/federal-rules-civil-procedure.

van Elswyk, Peter. 2018. "Un/qualified declaratives." PhD diss., Rutgers University - School of Graduate Studies. https://doi.org/10.7282/t3-asrd-fg29.

———. 2021. "Representing Knowledge." *The Philosophical Review* 130 (1): 97–143. https://doi.org/10.1215/00318108-8699695.

———. 2023. "Hedged testimony." *Noûs* 57 (2): 341–369. https://doi.org/10.1111/nous.12411.

———. 2024. "Hedging in discourse." *Synthese* 204 (3): 98. https://doi.org/10.1007/s11229-024-04733-6.

van Elswyk, Peter and Sapir, Yasha. 2021. "Hedging and the ignorance norm on inquiry." *Synthese* 199 (3): 5837–5859. https://doi.org/10.1007/s11229-021-03048-0.

van Elswyk, Peter and Willard-Kyle, Christopher. Forthcoming. "Hedging and the Norm of Belief." *Australasian Journal of Philosophy*.

Van Oosterum, Kyle. 2025. "Confucian Harmony, Civility, and Echo Chambers." *Journal of Applied Philosophy*, 1–23. https://doi.org/https://doi.org/10.1111/japp.12791.

Vermaire, Matthew. 2024. "Judgment's aimless heart." *Noûs*, 1–19. https://doi.org/10.1111/nous.12497.

Vosoughi, Soroush, Roy, Deb, and Aral, Sinan. 2018. "The spread of true and false news online." *science* 359 (6380): 1146–1151. https://doi.org/10.1126/science.aap9559.

Wang, Lu, Beauchamp, Nick, Shugars, Sarah, and Qin, Kechen. 2017. "Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes." *Transactions of the Association for Computational Linguistics* 5:219–232. https://doi.org/10.1162/tacl_a_00057.

Wang, Yuxi, McKee, Martin, Torbica, Aleksandra, and Stuckler, David. 2019. "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media." *Social Science & Medicine* 240:112552. https://doi.org/10.1016/j.socscimed.2019.112552.

Webber, Jonathan. 2013. "Liar!" *Analysis* 73 (4): 651–659. https://doi.org/10.1093/analys/ant081.

Whitmarsh, Lorraine and Corner, Adam. 2017. "Tools for a new climate conversation: A mixed-methods study of language for public engagement across the political spectrum." *Global Environmental Change* 42:122–135. https://doi.org/10.1016/j.gloenvcha.2016.12.008.

Willard-Kyle, Christopher. 2020. "Being in a Position to Know is the Norm of Assertion." *Pacific Philosophical Quarterly* 101 (2): 328–352. https://doi.org/10.1111/papq.12305.

Williamson, Timothy. 1996. "Knowing and asserting." *The Philosophical Review* 105 (4): 489–523.

———. 2000. *Knowledge and its Limits*. New York: Oxford University Press.

———. 2020. *Philosophical Method: A Very Short Introduction*. Oxford University Press.

———. n.d. "Justifications, Excuses, and Sceptical Scenarios." In *The New Evil Demon*, edited by Fabian Dorsch and Julien Dutant. Oxford: Oxford University Press.

Wood, Michael J. 2015. "Some Dare Call It Conspiracy: Labeling Something a Conspiracy Theory Does Not Reduce Belief in It." *Political Psychology* 37 (5): 695–705. https://doi.org/10.1111/pops.12285.

Worsnip, Alex. 2021. "The skeptic and the climate change skeptic." In *The Routledge Handbook of Political Epistemology*, 1st ed., edited by Michael Hannon and Jeroen de Ridder, 469–479. Routledge. https://doi.org/10.4324/9780429326769-55.

Zagzebski, Linda Trinkaus. 2012. *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford University Press.